AD_____

Award Number: DAMD17-98-2-8003

TITLE: Massachusetts Institute of Technology Consortium Agreement

PRINCIPAL INVESTIGATOR: Haruhiko H. Asada, Ph.D.

CONTRACTING ORGANIZATION: Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

REPORT DATE: March 1999

TYPE OF REPORT: Final II of Phase 2

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

19990901 036

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1999	Final (31 Dec 97 - 31 Dec 98)

4. TITLE AND SUBTITLE
Massachusetts Institute of Technology Consortium Agreement

5. FUNDING NUMBERS
DAMD17-98-2-8003

6. AUTHOR(S)
Haruhiko H. Asada, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

E-Mail:

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

14. SUBJECT TERMS

15. NUMBER OF PAGES
331

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

_____
PI - Signature   Date

Introduction

In this, our second progress report of the Phase Two Home Automation and Healthcare Consortium at the Brit and Alex d'Arbeloff Laboratory for Information Systems and Technology, M.I.T., we discuss all new findings, concepts, designs, and experiments obtained or accomplished in the past six months. Covered here are the diverse fields of home automation and healthcare research, ranging from human modeling, patient monitoring, and diagnosis to new sensors and actuators, physical aids, human-machine interface and home automation infrastructure. These results will be presented at the upcoming General Assembly of the Consortium held on October 27 - October 30, 1998 at MIT.

To better disseminate our technology to each sponsor company, this report in its entirety is posted at our web site: http://darbelofflab.mit.edu. Please visit our homepage and look at the consortium report section. The report section is password protected, which can be obtained from the representative of each sponsor.

Patentable ideas and inventions reported here have been filed as provisional patent applications. All sponsors will be notified of these provisional applications through Benjamin Palleiko, benp@mit.edu, of the MIT Technology Licensing Office. Those interested in pursuing the possibility of using the technologies should contact the Technology Licensing Office within six months after signing a non-disclosure agreement.

We trust that this report and the General Assembly will provide useful information for your field of research and development as well as for your business.

Thank you for your sponsorship that enables us to conduct our exciting research. We are looking forward to seeing you soon.

H. Harry Asada
Principal Investigator
Ford Professor and Director
d'Arbeloff Laboratory

Table of Contents

B

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Human Physiological Modeling

CHAPTER 1

Hemodynamic Modeling and State Estimation for Assessment of Cardiovascular
Health

R. Kamm, Y. Huang

Hemodynamic Modeling and State Estimation for Assessment of Cardiovascular Health

Roger D. Kamm and Yaqi Huang

1. Introduction

Noninvasive cardiovascular assessment in the home is currently limited primarily to the simple measurements of blood pressure and heart rate. The potential exists to monitor ECG as well, but few devices are capable either of continuous monitoring or of data interpretation beyond the obvious. Yet, there exists enormous potential in the subsequent analysis of these measurements that could potentially lead to more comprehensive, and more useful, information concerning the cardiovascular state of the individual.

Continuous measurement of blood pressure is now a reality with the recent development of systems that can be worn, either on the wrist or, in the case of the Ring Sensor, on the subject's finger. Miniaturized sensors and on-board electronics enable the device to convert the measured signal to a form more easily transmitted to a central computer for further processing and analysis.

The processing of this information is designed to extract all the useful information contained in the signal. In the case of the blood pressure pulse, clinicians have known, and made use of the fact that various aspects of the waveform contain information about the sate of the heart or the peripheral vascular network. For example, the maximum rate of pressure rise at the beginning of systole is indicative of the strength of cardiac contraction while the rate of decay of pressure during end diastole is a measure of peripheral vascular resistance, both of which are important parameters used in cardiovascular diagnoses.

The inference of cardiac parameters from peripheral measurements, however, is complicated by the changes in pulse shape that occur as the pressure wave propagates through the intervening arterial tree. Others have sought to overcome this problem by establishing the transfer function that describes the change in shape of the pulse between the aortic root and the peripheral measurement site. This method suffers, though, from the need for periodic calibration requiring arterial catheterization. An alternative approach, developed by our group, involves the use of a comprehensive model of the entire arterial system and heart. This computational model is used to create a *solution library* consisting of an extensive collection of peripheral pressure traces, each corresponding to a different set of system parameters, covering the entire range of possible parameter values. The solution library is further condensed by a two-step

1

process. First, each curve is represented by some small number of features (*feature extraction*). These features are selected so that they describe the shape and magnitude of the pressure waveform, and correspond in number to the set of *critical parameters*, those we seek to predict by our parameter estimation technique. Second, the dependence of each feature on the critical parameters is viewed as an *N*-dimensional surface and is mathematically represented by a *surrogate function*. The surrogate function itself is represented by a set of coefficients that are stored for later use in the parameter estimation procedure.

Parameter estimation begins with the measurement of arterial pressure by one of several non-invasive methods. The measured trace is processed in the identical manner as the computed waveforms to extract the features. An initial seed is chosen (a particular point in parameter-space) and a measure of the relative error between the features calculated from the measurement and the features corresponding to the initial seed point. Beginning at this point, a minimization routine is used to march down the error surface to eventually identify the point in parameter-space having the smallest error and therefore corresponding to the set of parameters that most closely match those of the subject from whom the measurements are taken.

Since the last progress report, our efforts have focused on several topics:

1) Methods to improve the computational efficiency of the model.
2) Development of a second model that emphasizes the cardiovascular control sytem.
3) Evaluation of our parameter estimation procedures.

Our progress in each of these areas is summarized below.

2. The Computational Model

The details of our model of the arterial system were given in the last progress report (see Report No. 2-1). Briefly, the model consists of a distributed representation of the entire arterial tree, extending out to the small peripheral vessels. Flow and pressure wave propagation is represented by the appropriate one-dimensional, non-linear partial differential equations for flow through a compliant network of vessels. Extensive validation studies have confirmed that all essential features of the pressure pulse are accurately portrayed by the model. Also included is a model for the left heart in which the elastic and viscoelastic behaviors are both represented.

In its then current form, the computational model required approximately 45 minutes to compute the pressure waveforms corresponding to each set of parameter values using a relatively powerful workstation. Since the accuracy of our parameter estimation routines can be improved by increasing the number of simulations contained in our library, we have looked into ways to reduce this time. Toward this end, a new method has been introduced to compute the time-dependent wall shear stress acting on the arterial wall.

Our previous approach was both memory-intensive and time-intensive. A new approach has now been introduced that is now in the final stages of testing which saves on both. In this approach, wall shear stress is computed using an expression involving only a single time-derivative as opposed to the previous method that required calculation of a convolution integral over many computational time steps. Our initial results indicate that there is no perceptible change in the solution, and a considerable savings in computational time.

3. Lumped Parameter Model with Cardiovascular Control

In certain situations, it is advantageous to use a model that accounts for the body's natural ability to compensate for changes in the environment. For example, a change in heart rate, contractility, or cardiac output might either reflect a pathological change or simply the normal response to an every-day maneuver such as standing from a seated or lying position. Clearly, any home diagnostic system would need to have the capability of distinguishing between these two potentialities. To develop this capability, we have begun to create a model of the cardiovascular control system. This model, while under development, is coupled to a simpler, and less computationally-intensive model of the arterial tree. In addition, the entire vascular network is represented, including the pulmonary and venous circulations.

Our starting point in this process is a model produced by Davis (1991) that has been regularly used as a teaching aid in courses on cardiovascular physiology and pathophysiology at MIT and the Harvard Medical School. The model represents left and right ventricles, systemic arteries and veins, and pulmonary arteries and veins as linearly expansile compartments connected by linearly resistive conduits (see Figure 1). The pumping action of the ventricles is modeled by time-varying ventricular compliances; the time variation is taken from canine experiments done by Suga et al. [1974, 1980]. The atria are not explicitly modeled: their effect is partially absorbed in the constituent parameters for neighboring compartments. The inertial effects of blood flow are ignored. A set of six first order time-varying differential equations describes the basic six-compartment system. A numerical simulation using a fourth-order Runge-Kutta integration method with adaptive time steps provides pressures at each of the six nodes of the circuit as functions of time. Volumes and flow rates are easily calculated from these pressures and the corresponding constituent relationships. Simulated data (e.g., pressures, flows, volumes) may be displayed dynamically as functions of time, or one variable may be displayed versus another (e.g., pressure-volume loops).
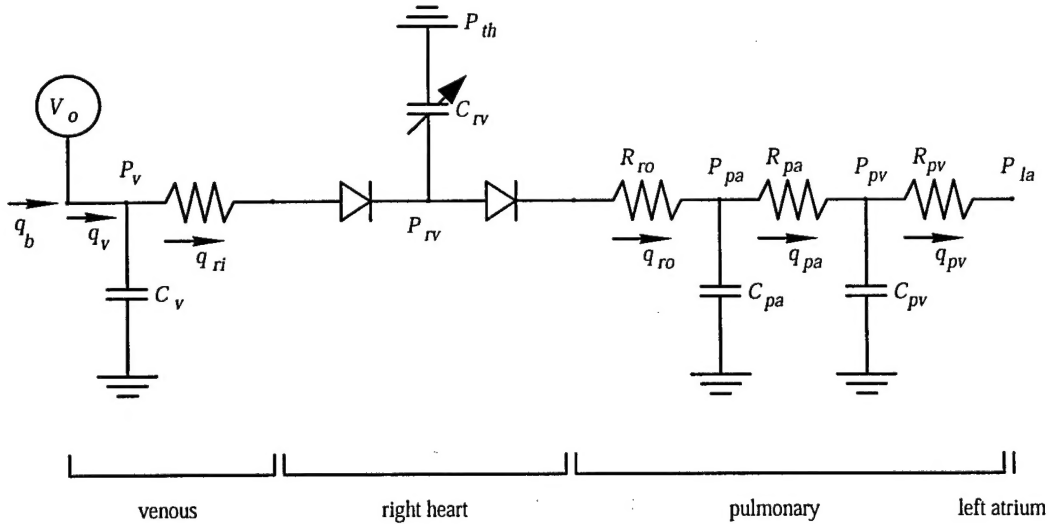
Figure 1. Electrical circuit analog of the venous and pulmonary circulations.

Control System. A first-order model of the baroreceptor reflex was implemented which controls heart rate, cardiac contractility, peripheral resistance, and venous blood volume to maintain a constant arterial blood pressure (Figure 2). Arterial baroreceptors transmit a signal to the autonomic nervous system (ANS) which is related to arterial blood pressure fluctuations. The ANS exerts homeostatic controlling actions on the hemodynamic system. A fast parasympathetic reflex arc via the vagus nerve tightly controls sino-atrial (SA) node firing rate, while slower sympathetic fibers modulate the strength of ventricular contraction and peripheral vascular tone, in addition to providing a slow contribution to heart rate. Four limbs of the baroreceptor reflex are modeled in the simulator: heart rate, cardiac contractility, peripheral resistance, and venous tone. The input to the controller is an "effective blood pressure deviation" measure which accounts for the threshold and saturation behavior of the baroreceptor [DeBoer et al, 1987].

$$dP_a(t)/dt = 18 \tan^{-1}\{(P_a(t) - P_{SP})/18\}$$

where $P_a(t)$ is the arterial pressure, and P_{SP} is the set-point pressure. The arctangent mapping preserves the baroreceptor sensitivity around the set-point, while limiting the maximum signal to about 28 mmHg deviation from the set-point.

The controller is modeled as a set of four parallel linear filters, each of which produces an output that modulates the value of a controlled parameter. The time course of each limb's reflex response is defined by its impulse response. This is convolved with $dP_a(t)/dt$ to generate a reflex signal, which is then used to proportionally modify the appropriate parameter. For example, the peripheral resistance, R_a, would be given by:

$$R_a(t) = R_{a0} - \sum_{k=0}^{J} b(k)p'_a(t-k)dk$$

4

where R_{a0} is the reference resistance, and b(k) is the impulse response of the "resistance" filter which incorporates the appropriate delay time and gain. The time step is 0.5 seconds, so the convolution is done over a 30-second period. Parameters for the model were determined from published findings on open-loop gain, and time delays associated with receptor, nerve, and effector organ response.

Timing was obtained from Berger, Saul and Cohen (1989). Their study measured the impulse responses for sympathetic and parasympathetic baroreceptor reflexes in man, and this data was used to design our model's filters (figure 3).
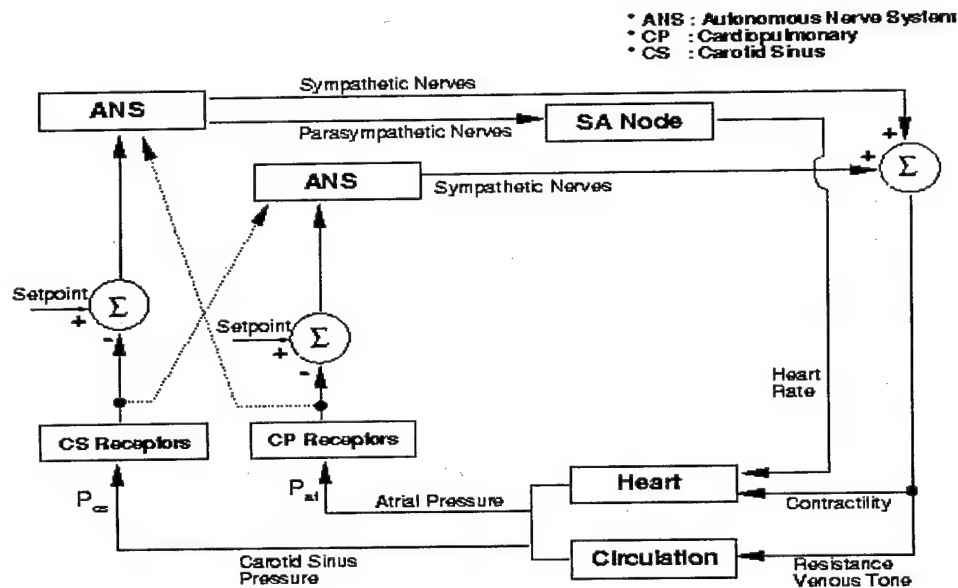

Figure 2. The control system loop. Adapted from Davis , 1991.

The filter for each limb of the reflex is scaled to the overall filter gain, or reflex sensitivity. The gain values used in the model are shown in Table 1.

Reflex Limb	Gain	Nerve Timing
RR interval	18 ms/mmHg	Beta symp & parasymp
LV contractility, $C_{L,sys}$	0.007 ml/mmHg/mmHg	Beta symp
RV contractility, $C_{R,sys}$	0.021 ml/mmHg/mmHg	Beta symp
Resistance, R_a	0.011 PRU/mmHg	Alpha symp
Venous zero-p volume	26.5 ml/mmHg	Alpha symp

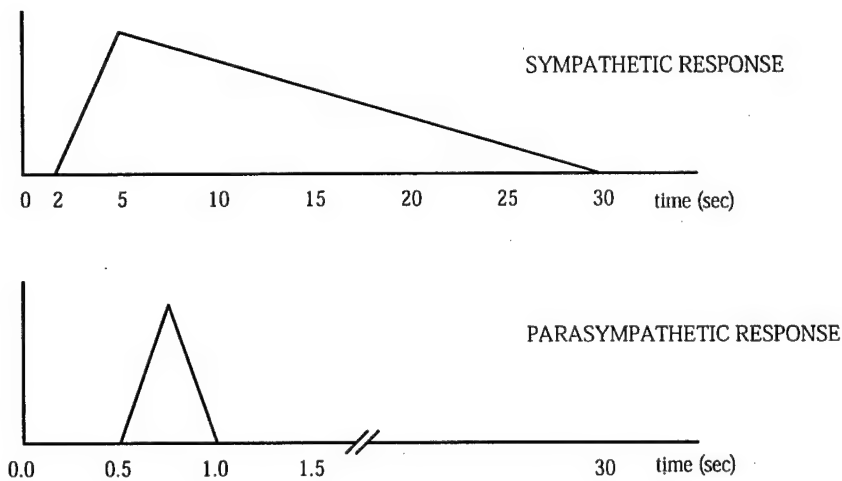Table 1. The reflex gain values used in the simulator.

Figure 3. Sympathetic and parasympathetic feedback impulse responses for the baroreflex control model. The rapid parasympathetic response begins in 0.5 seconds and is complete in 1.0 second. The slower sympathetic response begins after a latency of 2 seconds, peaks at 5 seconds, and lasts 30 seconds.

The baroreflex heart rate and resistance sensitivities were taken from Deboer (1987). Observations on the variability of contractility were used to set the gain of the APB-to-contractility limb. A reasonable range for venous blood volume deviations was inferred from Guyton's (1986) comments on postural variability of venous volume. Although direct human experimental data are lacking for some of the feedback control parameters, the control system implemented in the model was successful in achieving the objectives of a first-order approximation to the physiologic system. The closed loop gain, temporal and oscillatory behaviors are quite similar, but not identical, to those observed in humans (Davis, 1991).

More recently, this model has been adapted for use in an expert system designed to identify the model parameters from a set of output data which would be obsevable in the ICU setting.

The block diagram of our enhanced reflex control system is shown in figure 3.1. Both reflex systems, cardio--pulmonary and carotid—sinus baroreflex, are shown with their respective effector mechanisms: contractility, heart rate, venous tone, and vascular resistance. The parameters of the model were based on literature values. The gains for the reflex limbs can be specified independently (see Appendix B for pictures of the graphical user interface) which makes it a powerful tool for testing of specific hypotheses involving alterations of the reflex system.

Known interactions between the carotid-sinus baroreflex and the cardio-pulmonary reflex (shown as dotted lines) are now being implemented in the model. The entire reflex system is to be finished later this year.

Hemodynamic System.

Carotid Transmural Pressure. To take into account the effects of gravity induced changes in carotid sinus transmural pressure we made the sensed pressure of the carotid sinus reflex a function of orientation in the gravitational field:

$$P_S = P_{CS} - \rho gh \cdot \sin (\alpha - \pi/2)$$

where ρgh is of the order of 10 mmHg and α represents the orientation of the carotid artery with respect to the gravitational field.

Venous Transmural Pressure. The reference pressure acting externally on the veins is changed to produce the tendency for venous pooling when the legs are gravitationally dependent. This effect can be implemented in the source code with any specified function of time.

Capillary Filtration. Gravity induced capillary filtration is modeled as an overall reduction in total blood volume. Total blood volume is a system parameter that can be specified by the user through the graphical user interface (GUI). Furthermore, it is possible to specify total blood volume in the source code as a function of time to associate fluid loss into the interstitium with a certain intervention. Functional forms for blood loss is taken from the literature.

Standing or Tilt-Table Experiments. To initially test the performance of this model under conditions that mimic those that might be experienced at home, we used data obtained from studies in which the subjects went from supine to erect, either on a tilt table or under their own power. For this purpose, we implemented:

- changes in hydrostatic pressure at the carotid sinus baroreceptor due to changes in posture.

- a rapid change in venous transmural pressure which leads to blood pooling in the veins.

- changes in total blood volume due to an increased rate in capillary filtration.

We compare our simulation results to two experimental studies that specifically looked at the short-term circulatory response to changes in posture or the application of lower-body negative pressure. Brauer and co-workers looked at the beat-to-beat variations in heart rate as a function of tilt-speed. We simulated their experimental protocol and found that the temporal behavior of the heart rate response simulation agrees well with the experimental finding. The heart rate initially increased by 15 beats/min in the simulation (25 beats/min in the measurement) after which the rate fell to a level corresponding to a sustained increase of 7 beats/ min (5 beats/min in the measurement).

Sprangers et al. investigated the changes in heart rate and blood pressure at the onset of tilt and an active change in posture (angle of tilt: 90 degrees in 3 sec). The experimental results show that an active change in posture actually calls for a more dramatic transient change in heart rate and blood pressure than a passive one. The time course and magnitude of the transient behavior of our simulations seem to mimic the data on standing.

4. Parameter Estimation

Viewed in the simplest terms, our objective is to extract from a measured waveform, indices that characterize cardiovascular health. Our work to date has focused on three: systemic vascular resistance (SVR), left ventricular contractility (Elv,max), and end diastolic volume (EDV). SVR is a significant factor in subjects with high blood pressure and determines the afterload felt by the heart. Contractility is a useful measure of the heart's ability to pump. A rise in EDV is indicative of progressive degeneration in congestive heart failure. These three, in addition to mean, systolic, and diastolic pressure, provide a good basis for clinical evaluation, and also exert a significant influence over the shape of the pressure waveform. A fourth index, systolic fraction, is also predicted by the method, but is of little diagnostic value.

The features we have used to characterize these four indices are those commonly used by cardiologists, and include: mean pressure, pulse amplitude, maximum systolic dp/dt, and diastolic dp/dt. In addition, we have examined the possibility of augmenting the pressure wave measurement with measurements of the volume flow rate trace. Three locations have been evaluated, the carotid, radial and brachial arteries.

Our procedures for parameter estimation are tested initially using computer-generated test cases. Here we present two types of simulation, those corresponding to actual points in the solution library, and those corresponding to arbitrary locations in parameter space. The former provides us with some indication of the optimization method's accuracy if we

were to densely populate parameter-space with computed solutions. The latter can help us to determine whether the current library is sufficient. Results are presented below for one index, SVR (note that SVR is a normalized quantity in these). The errors are seen to be quite reasonable, on the order of several percent. Errors in the other parameters are somewhat larger, but still average only about 6% for EDV and 11% for Elv,max when the carotid artery is used as the measurement site.

Table 2. SVR at library points.
Locations are identified following Table 2.

library point #		25	37	50	85
real SVR value:		0.9999	0	0.3333	0.6666
errors (%)					
at each location	1	0.6339	5.6508	0.42	4.75
	2	0.0044	0.6280	6.28	3.98
	3	0.1333	0.0045	9.02	3.44
	4	0.0031	0.0328	0.19	0.28
	5	4.0012	0.0906	0.85	0.82
	6	4.9413	1.0891	0.08	4.69
mean error (%)		1.62	1.25	2.80	3.00

Table 3. SVR at random test points

test point #		1	2	3	4
real SVR value:		0.5667	0.8467	0.3133	0.4553
errors (%)					
at each location	1	0.47	7.07	0.3740	3.1086
	2	8.81	16.24	4.4179	10.9003
	3	5.01	19.90	4.7503	12.7824
	4	0.47	0.61	1.5244	0.6519
	5	0.54	0.01	0.5564	0.9545
	6	0.70	0.27	0.1404	0.1261
mean error (%)		2.67	7.35	1.96	4.75

Alternative methods based on a description of the pressure waveform using orthogonal functions, either Fourier series or wavelet representations, are also being developed. While these are less intuitive to the cardiologist, they are better descriptors of the waveform from a mathematical perspective. In addition, by using one of these orthogonal representations, it is easier to discern the information content of the signal.

The practical import of this is that the number of parameters that can be estimated should not exceed the minimum number of features needed to completely describe the waveform.

With any of these methods of signal representation, several potential problems will need to be overcome. In particular, if the surface in N-space representing the feature dependence on the parameters has multiple minima, then the possibility exists of getting "trapped" at a local minimum rather than reaching the global minimum. One example of this can be seen in Fig. 4 showing the dependence of the four current features on SVR when the other three parameters are held constant. In general, however, the curves are monotonic and can be adequately represented by a low-order polynomial.
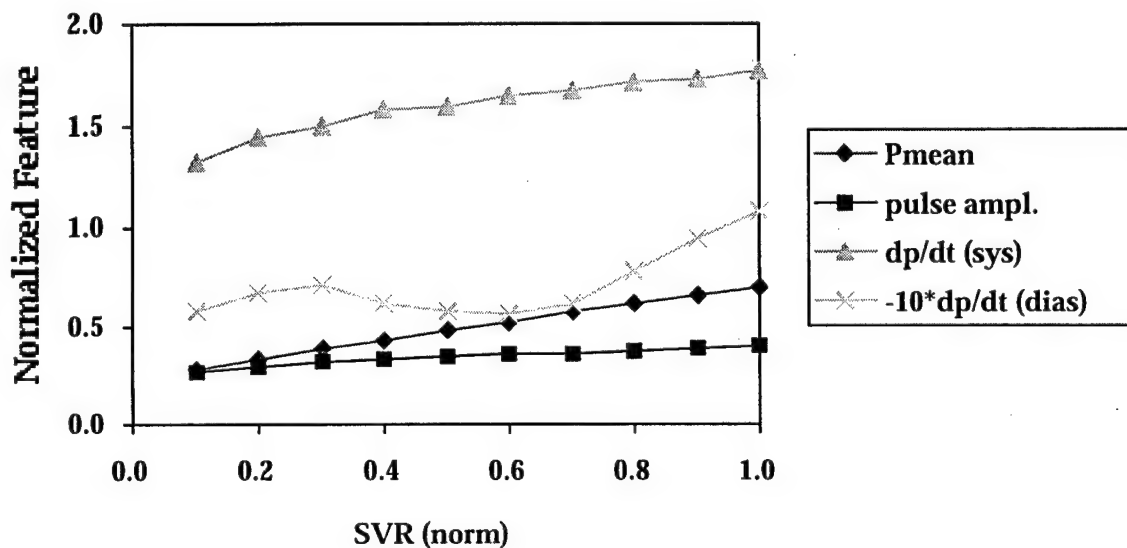

Figure 4. The dependence of each of the four cardiovascular features on Systemic Vascular Resistance. Note the two minima in the curve for $(dp/dt)_{dias}$.

Literature Citations

Berger, R.D., Saul, J.P., Cohen, R.J. Transfer function analysis of autonomic regulation I. Canine atrial rate response. *American Journal of Physiology*, 256 (Heart Circulation Physiology 25): H142-52, 1989.

Blomqvist, G.C., et al., Handbook of Physiology (eds.: Shepherd, T.J. and Abboud, F.M.), section 2 (The Cardiovascular System), vol III (Peripheral Circulation), part 2, 1983.

Brauer, G., et al. Zum Verhalten der Herzfrequenz des Menschen bei unterschiedlicher Geschwindigkeit des "Ubergangs vom Liegen zur Kopf"artsposition, Acta biol. med. germ. 34: 1153--1157, 1975.

Davis, T.L. Teaching Physiology Through Interactive Simulation of Hemodynamics. M.S. thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology: February, 1991.

DeBoer, R.W., J.M. Karemaker, and J. Strackee. Hemodynamic fluctuations and baroreflex sensitivity in humans: a beat-to-beat model. Am. J. Physiol. 253 (Heart Circ. Physiol. 22): H680-H689, 1987.

Guyton, A.C., T.G. Coleman, R.D. Manning, Jr., and J.E. Hall. Some problems and solutions for modelling overall cardiovascular regulation. Math. Biosci. 72: 141-155, 1984.

Mark, A.L. et al. Cardiopulmonary baroreflex in humans, In: Handbook of Physiology. The Cardiovascular System. Peripheral Circulation and Organ Blood Flow. Bethesda, MD: Am. Physiol. Soc., 1983, Sect. 2,
Vol. 3, Chap. 21, pp. 795-814.

Sprangers, R.L.H., et al. Initial Circulatory Responses To Change in Posture: Influence Of The Angle And Speed Of Tilt, Clinical Physiol. 11: 211--220, 1991.

Suga, H. and Sagawa, K. Instantaneous Pressure-Volume Relationships and Their Ratio in the Excised, Supported Canine Left Ventricle. *Circulation Research*, Vol 35, July 1974.

Suga, H., Sagawa, K., Demer, L. Determinants of Instantaneous Pressure in Canine Left Ventricle: Time and Volume Specification. *Circ. Res.* 46: 256-263, 1980.

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Patient Monitoring and Diagnosis

CHAPTER 2

Ring Sensor Project: New Developments on Noise Reduction and Miniaturization
B-H Yang, K-W Chang, S. Rhee, Y. Zhang, H. Asada

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Ring Sensor Project: New Developments on Noise Reduction and Miniaturization

Boo-Ho Yang, Sokwoo Rhee, Yi Zhang
Kuowei Chang and Harry Asada

1 Introduction

Ambulatory monitoring is a key technology for high-quality home healthcare. In the ring sensor project, we have developed a twenty-four hour patient monitoring system using a compact, wearable sensor in a ring configuration. Many prototype ring sensors have been developed to prove the concept of the ring sensor and exhibit the usefulness of the ambulatory sensor system.

One of the main issues is how to reduce signal noises. It is inherent for a wearable sensor such as the ring sensor to undergo a lot of noises. In fact, we often found in the ring sensor that the photoplethysmograms transmitted from the ring sensor are corrupted with noises in a significant degree. Since quality of twenty-four hour patient monitoring systems depends on consistent accuracy of the sensor readings, it is important to develop an efficient algorithm to reduce the noise effect from sensor. In this report, we present a new noise protection method based on the auto-correlation functions of the signals. It will be shown that, with this new technology, the heart rate of the patient can be consistently and accurately monitored even when the original sensor signals are significantly damaged by a noise.

Another main issues in developing the ring sensor is how to efficiently design a miniaturized ring sensor. Miniaturization of the sensor is extremely important for successful ambulatory monitoring. In the last progress report, we presented the development of a prototype miniaturized ring sensor. The ring sensor has a single PC board on the top of the ring and all the electrical functionalities including analog signal conditioning, digital processing and wireless transmission were mounted on the board. From the experience of the development, we found that it would give more robustness against electrical uncertainty while keeping the overall size compact if we separate some of the functionalities in two PC boards. Based on the consideration, we developed a new prototype miniaturized ring sensor. Details of the design of the ring sensor will be presented in this report.

1

2 Noise Reduction using Correlation Functions

2.1 Introduction

In all kinds of wearable sensors, one of the most difficult problems to deal with is the noise issue, which includes elimination of ambient light and motion artifact reduction. In our ring device, this issue is very critical due to the fact that optical sensors are used to detect an extremely faint signal which is buried in relatively strong noise from many sources. Unlike other types of sensors, optical sensors are very susceptible to all kinds of disturbances that affect the path of the photons. For example, the air gab between the sensor (photodiode in this case) and the skin gives a drastic influence on the quality of the received signal. Even a faint trembling of the hand can change the air gab and the obtained signal might swing irrespective of the actual heart beat. Even changing the intensity of ambient light can suddenly increase or decrease the amount of photons received by the optical detector.

Although there are thousands of reasons that can disturb the optical path of the ring device, most of them belong to two categories; motion artifact and ambient light change. Motion artifact is mainly a relatively low frequency disturbance, and it is essentially non-periodic as long as it is not caused by a intentional periodic movement. The change of ambient light is also non-periodic in a normal room condition. These two major noise factors can be effectively treated by a combination of classical low pass filter and signal correlation technique.

If the noise is extremely stronger than the pulse signal, there is very little chance to reconstruct the original waveform no matter what kind of fantastic signal processing method is used. For example, if the noise is so strong that the detected signal goes beyond the range of detection (i.e. the signal becomes saturated), it is impossible to recover the pulse signal which is completely buried in the noise. But if we limit the scope of noise reduction issue in the ring sensor as a 24-hour pulse rate monitoring device, then the correlation method is one of the strongest method that can be used. In other words, the correlation method can be best used for the continuous monitoring of the pulse rate even in the existence of severe noise sources, since this method works surprisingly well in separating the periodic component of pulsation of the heart and the non-periodic noise components.

In this report, a theoretical analysis of applying the correlation technique is first explained. Secondly, a numerical simulation is done to verify usefulness of correlation method in detecting a desired signal with a low signal-to-noise ratio. Finally, several signal processing results with real experimental data are presented.

2.2 General Description of Signal Conditioning Process

The autocorrelation method is a very powerful tool in catching a periodic signal buried in a random signal. This method is especially useful in application on our ring sensor since the technique works even under relatively low signal-to-noise ratio. The noise sources that contaminate the heart beat signal are mostly non-periodic (ex: artifact by random motion) or periodic in much higher frequency than heart beat (ex: ambient light). So if we get rid of the high frequency noise by a low pass filter and deal with non-periodic low frequency noise using autocorrelation function, both kinds of the noise can be effectively reduced. The following Figure 1 shows the flow chart of the signal processing.
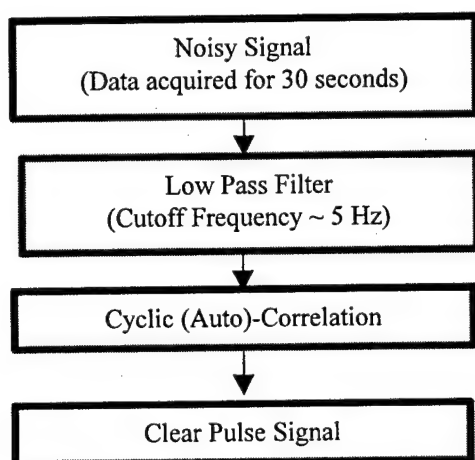
```
┌─────────────────────────────────┐
│         Noisy Signal            │
│  (Data acquired for 30 seconds) │
└─────────────────────────────────┘
                │
                ▼
┌─────────────────────────────────┐
│         Low Pass Filter         │
│   (Cutoff Frequency ~ 5 Hz)     │
└─────────────────────────────────┘
                │
                ▼
┌─────────────────────────────────┐
│      Cyclic (Auto)-Correlation  │
└─────────────────────────────────┘
                │
                ▼
┌─────────────────────────────────┐
│        Clear Pulse Signal       │
└─────────────────────────────────┘
```

Figure 1 : Signal Processing Flow Chart

As the frequency domain analysis is necessary, the continuous stream of signal detected by the ring sensor is sliced into a unit of several seconds. (In this case the unit is chosen to be 30 seconds.) This slice of signal passes through a low pass filter. This filter mainly removes the relatively high frequency noises higher than 5 Hz including 60 Hz frequency component of the room light. This low pass filter is implemented in the device itself in the form of hardware. We can add some extra low pass filter in the software side if necessary.

The low pass filtered signal goes through autocorrelation function. With this process, most of the non-periodic components are removed from the signal and only periodic components of relatively low frequency will survive, and most likely they are supposed to be heart beat.

Actually the "windowing" process (slicing the stream of signal) is one limiting factor of this signal processing in real time analysis since an abrupt change of the heart beat cannot be promptly detected. The correlation method works better with a larger window, and the response delay to the change of heart operation becomes shorter as the windows is shorter. This reveals a kind of trade-off that should be considered in designing the signal processing algorithm.

3

The autocorrelation function works as a kind of smoothing factor. It compares the multiple waveforms in a signal slice, and derives an average of the waveforms. Due to this kind of averaging effect, the sharp peaks of the waveforms often become blunt and sometimes even disappear after correlation. This is another drawback of this method. The resolution of signal analysis using autocorrelation method can be improved to some extent by shortening the length of the window. This ends up to another factor of trade-off.

2.1.1. Theoretical Description of the Autocorrelation Function

Autocorrelation method is an averaging function that appears in random signal processing area. In this section, a brief explanation of autocorrelation and its application on our ring sensor will be presented.

The time-average of a discrete random signal x[n] is defined as :

$$< x[n] > \equiv \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n \to -N}^{N} x[n] \tag{1}$$

The autocorrelation function is defined as follows,

$$R_x[k] \equiv < x[n]x[n-k] > = < x[n]x[n+k] > \tag{2}$$

The autocorrelation function of a periodic signal with period N is also periodic with period N :

$$\begin{aligned} &if \ \ x[n+N] = x[n], \\ &then \ \ R_x[k+N] = < x[n]x[n-k-N] > = < x[n]x[n-k] > = R_x[k] \end{aligned} \tag{3}$$

For a signal having no periodic component, the autocorrelation function with a relatively large lag k approaches the square of the mean :

$$\lim_{|k| \to \infty} R_x[k] = \mu_x^{\ 2} \quad if \ \ x[n] \ \ is \ \ non-periodic \tag{4}$$

This means that the autocorrelation function of a non-periodic signal doesn't contain AC component with a certain value of k of which absolute value is larger than zero, as long as the sampling frequency is high enough relative to the frequency of the signal. For example, if we sample with 1 kHz, and take data for 30 seconds (total data point is 30000), the autocorrelation of the non-periodic signal becomes almost constant with a relatively large value of k such as k=1000.

Now let's look at the crosscorrelation function. A crosscorrelation function is defined as :

$$R_{xy}[k] \equiv < x[n]y[n+k] > = < x[n-k]y[n] > \tag{5}$$

4

If x[n] and y[n] are two sinusoidal waves with different frequencies, the crosscorrelation function of those two signals is zero, which means that they are uncorrelated. On the other hand, the crosscorrelation of two sinusoidal waves having the same frequency is a sinusoidal wave at the same frequency.

Let's take two sinusoidal signals with zero means,

$$x[n] = A_x \cos(2\pi f_x n + \phi_x) \tag{6a}$$

$$y[n] = A_y \cos(2\pi f_y n + \phi_y) \tag{6b}$$

The crosscorrelation function is

$$
\begin{aligned}
R_{xy}[k] &= < x[n]y[n+k] > \\
&= A_x A_y < \cos(2\pi f_x n + \phi_x)\cos(2\pi f_y (n+k) + \phi_y) > \\
&= \frac{A_x A_y}{2} < \cos(2\pi (f_x + f_y)n + 2\pi f_y k + \phi_x + \phi_x) > \\
&\quad + \frac{A_x A_y}{2} < \cos(2\pi (f_x - f_y)n - 2\pi f_y k + \phi_x - \phi_x) >
\end{aligned}
\tag{7}
$$

If $f_x \neq f_y$, both of the term of (7) are sinusoidal functions of n, so that their mean are both zero, and $R_x[k]$ is also zero. If on the other hand, $f_x = f_y = f$, the second term is a function of k, not n, so that

$$R_{xy}[k] = \frac{A_x A_y}{2} < \cos(2\pi f_y k - \phi_x + \phi_x) > \tag{8}$$

which has the same frequency as the original signal (6a) and (6b). From this result, we can say that if y[n] is a non-periodic signal (which means the frequency is zero) with zero mean, the crosscorrelation of a periodic signal x[n] and y[n] becomes zero. If the signal y[n] is a non-periodic but not a zero mean, the crosscorrelation function becomes a constant. This means that the crosscorrelation of x[n] and y[n] doesn't have any AC component.

$$\textit{if } y[n] \textit{ is non-periodic, } R_{xy}[k] = const \tag{9}$$

Using this result of (8), we can also get the autocorrelation function of a periodic signal with an arbitrary waveform.

Let x[n] be a periodic signal expressed as a discrete Fourier series :

$$x[n] = \sum_{i=0}^{\infty} A_i \cos(2\pi i \frac{n}{N} + \phi_i) \tag{10}$$

5

From (8), we know that all the different frequency components are uncorrelated with each other. Therefore, we get the following result :

$$R_x[k] = \sum_{i=0}^{\infty} \frac{A_i}{2} \cos(2\pi i \frac{k}{N})$$ (11)

Equation (11) shows that autocorrelation of any periodic function has the same waveform and frequency as the original signal x[n].

The signal received by the ring sensor can be expressed as following :

$$x[n] = s[n] + d[n]$$ (12)

where s[n] is a periodic heart beat signal and d[n] is a non-periodic noise. The autocorrelation function of x[n] is :

$$R_x[k] =< x[n]x[n+k] >=< (s[n]+d[n])(s[n+k]+d[n+k]) >$$
$$=< s[n]s[n+k] > + < d[n]s[n+k] > + < s[n]d[n+k] > + < d[n]d[n+k] > \quad (13)$$
$$= R_s[k] + R_{sd}[k] + R_{ds}[k] + R_d[k]$$

From (4), we can see that $R_d[k] \approx$ constant, and $R_{sd}[k]$ and $R_{ds}[k]$ are also constants according to (9). From (11), $R_s[k]$ contains the waveform and the frequency of the original signal s[n], and this $R_s[k]$ is the only AC component in $R_x[k]$. In the ring sensor application, this means that only periodic heart beat signal survives and non-periodic noises are removed by applying autocorrelation.

$$R_x[k] = R_s[k] + const$$ (14)

2.3 Numerical Simulation

In order to test the plausibility of the autocorrelation method, an artificially-made periodic signal and a random noise were put into simulation. The periodic signal generated by software has the frequency of 1 Hz and the signal was sampled with 1 kHz sampling rate. (This is the s[n] in equation (12).) The total data acquisition time was 10 seconds. Figure 2 shows the periodic input signal.
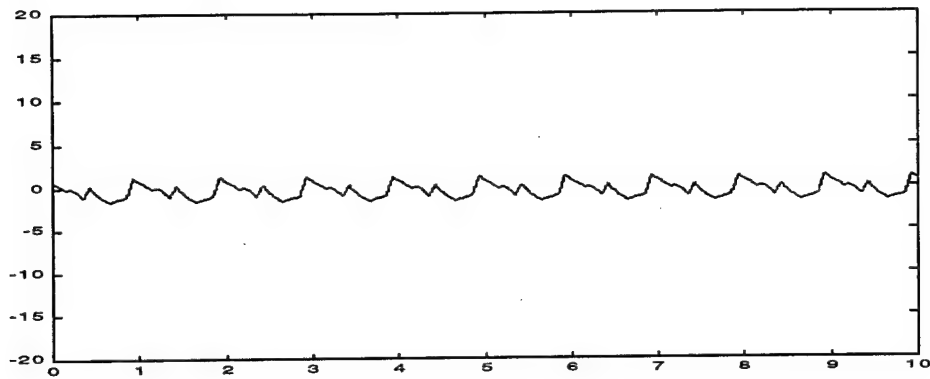
6

Figure 2 : Periodic Source with 1 Hz Frequency s[n]

The random noise is generated with much higher amplitude to comply with the real situation that has a low signal-to-noise ratio. This noise is not a "white noise." This noise is generated randomly without any consideration for frequency domain. If we see this noise in frequency domain, it also shows a random behavior. Figure 3 shows the random noise generated by a Matlab function. (This is the d[n] in equation (12).)
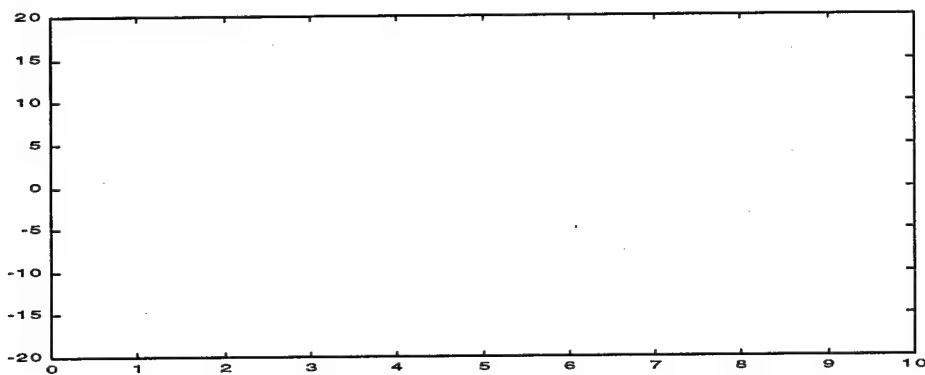

Figure 3 : Random Noise d[n]

Figure 4 shows the combined signal of Figure 2 and Figure 3. This represent the x[n] in equation (12). The s[n] signal (drawn as a white line for comparison) seems to be completely buried in the noise since the signal to noise ratio is low.
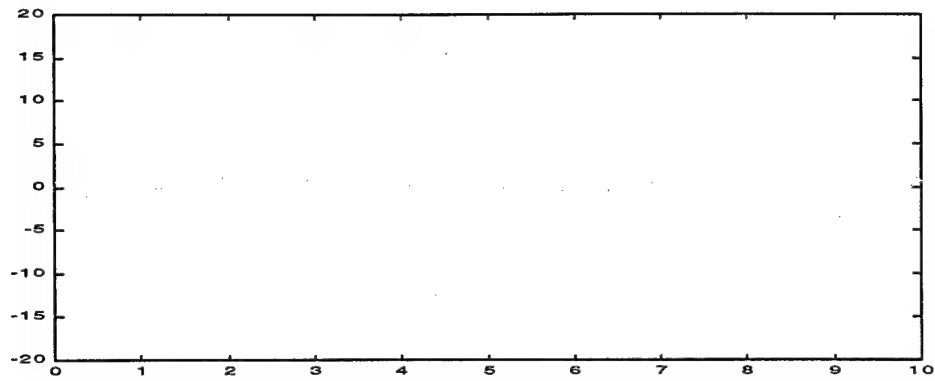
Figure 4 : Combined signal x[n]

The autocorrelation function of this combined signal is shown on Figure 5. As was expected by equation (13), we can see that the periodic component of s[n] was recovered, but the recovered signal is still not so clear. If we apply the autocorrelation twice, we can get a much better periodic signal. The result of the second autocorrelation function is shown on Figure 6.
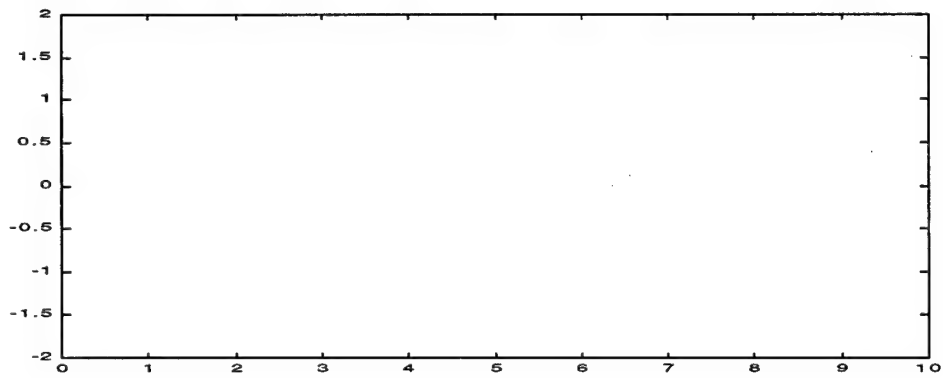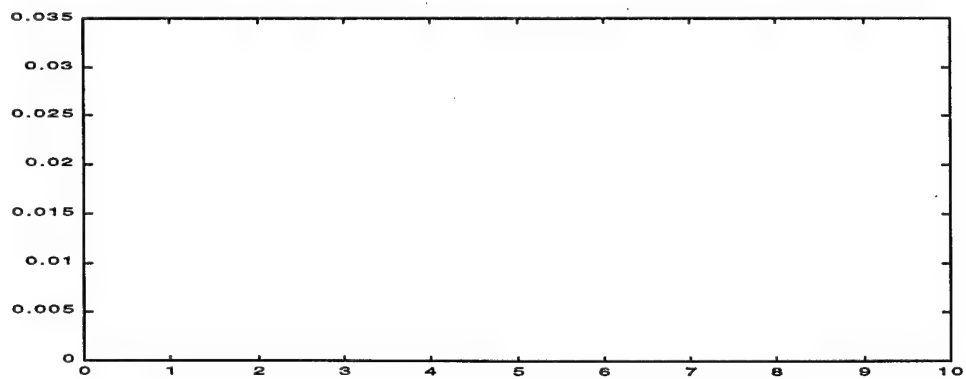

Figure 5 : Autocorrelation function of x[n]

Figure 6 : Second autocorrelation function of x[n]

8

After applying autocorrelation, we could recover the frequency of the periodic component exactly. The waveform was not completely recovered due to the low signal-to-noise ratio and the averaging effect as was stated before. But we can recover the periodic component to some extent. This means that we can at least recover the heart rate of the detected signal from a ring sensor.

2.4 Experiment

2.4.1 Experimental Setup

To support the theoretical result and the numerical simulation, an experimental setup was built. Figure 7 shows a schematic diagram of the setup. The signal obtained from the photodiode (detector) goes into a hardware signal processing unit. First, DC component in the signal is removed since we only need oscillating AC part, and a low pass filter is applied to get rid of high frequency noise.

The signal is then amplified by about a couple of thousands times, which raises the voltage around 2~3 volts that is in the detectable range of the A/D converter. The signal can be sampled with a rate of up to 20 kHz by using a custom-built windows NT-based real time data acquisition system. After the data become ready in the computer memory, they are tempered by the software-based signal processing techniques such as autocorrelation.
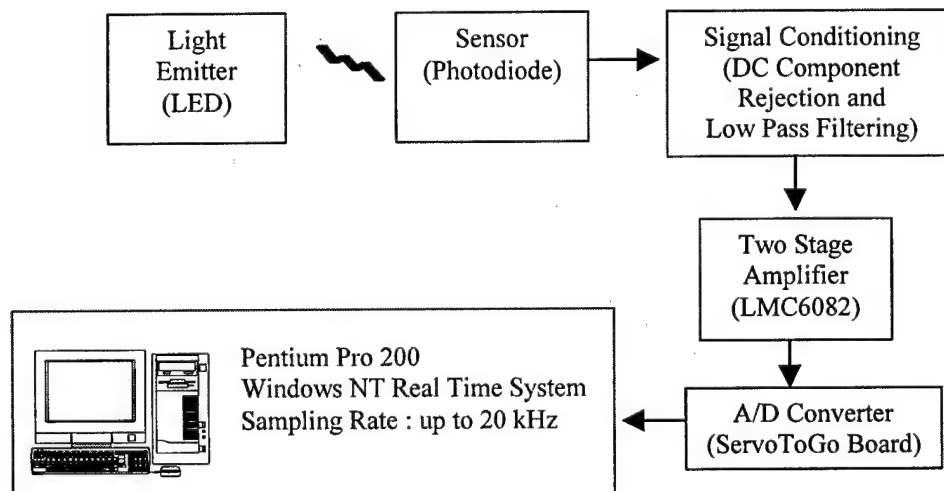

Figure 7 : Experimental Setup

2.4.2 Experiment Result

The simulation result was also approved by the real experimental result. Figure 8 shows the original signal detected by the ring sensor and the signal processed by the

9

autocorrelation technique. The original signal was obtained from the ring sensor when the hand was tightly clenched. As can be seen from the figure, the original signal was seriously distorted and contaminated with noise. The signal even shows the state of bad saturation because of the hand grip. But after the correlation technique was applied, the signal was arranged in a neat shape and the high frequency noise was also removed. Although the exact "pulse shape" could not be recovered, it can be seen that the periodic component of the detected signal was reconstructed by the autocorrelation method, which means that we can clearly estimate the pulse rate even in the worst situation.
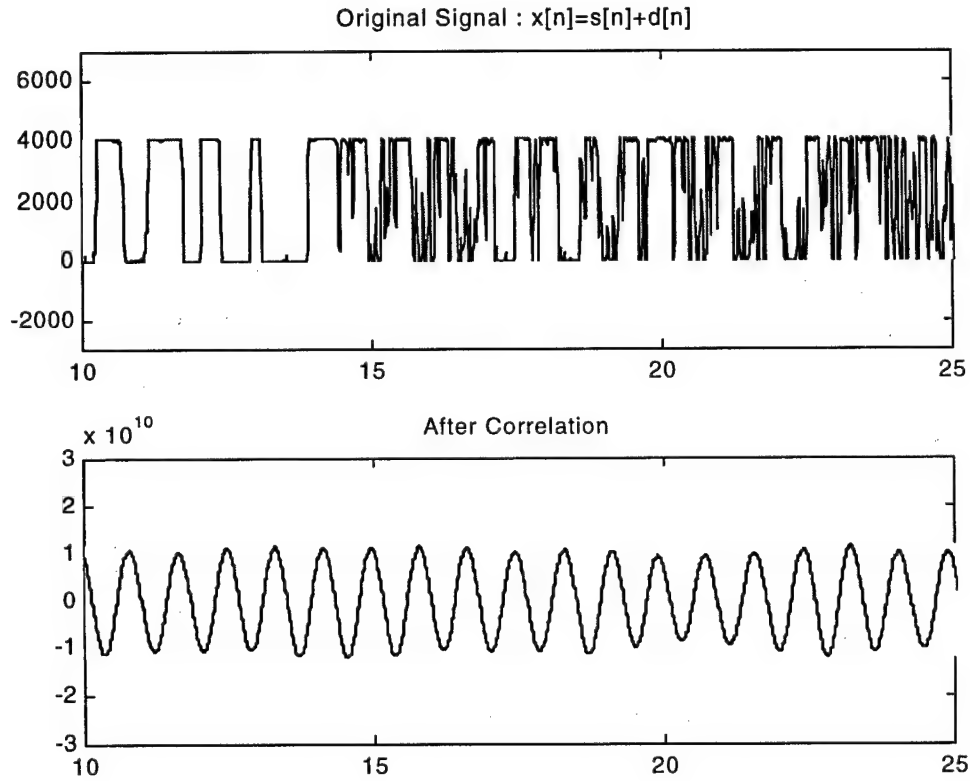

Figure 8 : Original ring sensor signal and the result after autocorrelation

This example does not explicitly show the second peak of the heart beat after autocorrelation since the periodic characteristic of the second peak in the original signal was not apparent and was seriously contaminated with many kinds of noise. The next example shows that even the second peak of the heart beat can be reconstructed as long as the original signal contains a clear behavior of the second peak as shown in Figure 9.
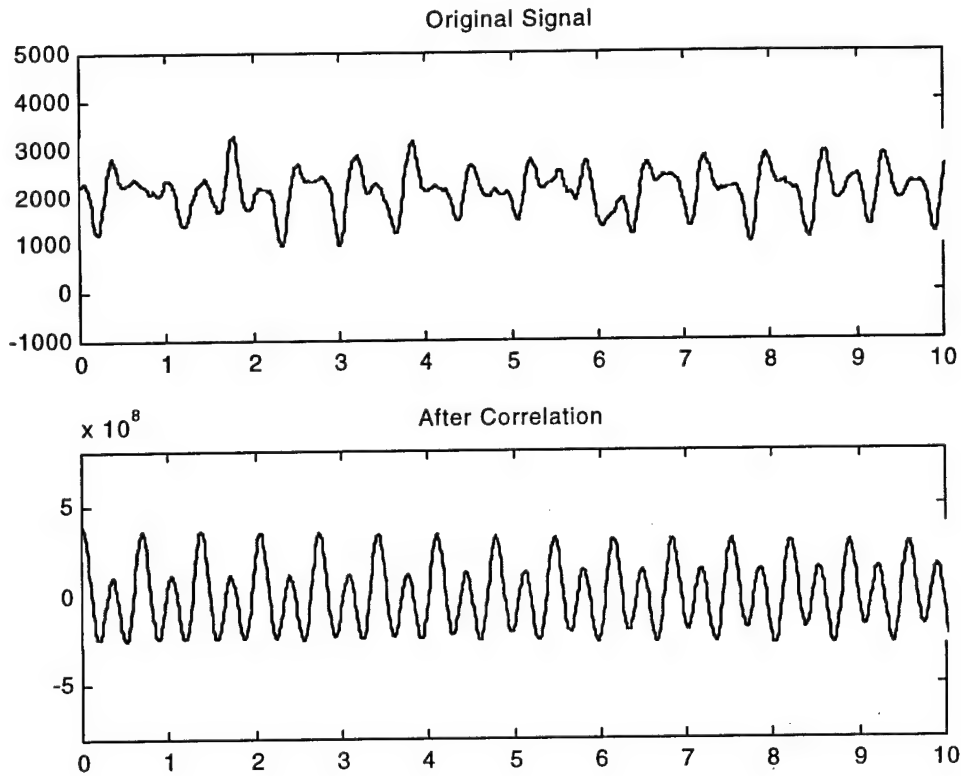
10

Figure 9 : Autocorrelation function showing the second peak of the heart beat

As we can see in the experimental results, the autocorrelation function method shows certain limitation in reconstructing the smaller peaks in the waveform. This is mainly due to the averaging effect of correlation. We can better reconstruct the waveform by using a shorter window for slicing the signal at the expense of effective noise reduction. If the noise is relatively less, then it is possible to reconstruct the original waveform with higher resolution. On the contrary, we have to stay with getting only the exact pulse rate in the case of severely low SNR. In our ring application, of which main emphasis lies in continuous heart beat monitoring, the pulse rate itself will have a great clinical importance.

2.5 Conclusion

A signal processing technique that can effectively remove the noise caused by motion artifact and ambient light was presented. Although the autocorrelation function technique has certain limitation in itself, it shows a generally good result in rejecting non-periodic noise when combined with classical noise reduction technique such as low pass filtering. A numerical simulation result was also presented to support this concept and it showed that this method can be used even with a low signal-to-noise ratio which is the case of our ring sensor. The experimental result also showed that the periodic component of the heart beat buried in the non-periodic noise can be effectively reconstructed by this method.

11

In the future, the input modulation method can be combined with correlation function technique to give a better noise reduction result. If we modulate the input signal with a certain waveform and a certain frequency, and do cross-correlation with the output signal, it is expected to give a more effective noise reduction result. In this case, the waveform of the heart beat can be better reconstructed by using a correct frequency and waveform in the input signal, hence improving the waveform reconstruction which is necessary to obtain further clinical information.

3 New Design For Miniaturization

3.1 Introduction

Health monitoring at home is highly demanded due to the increasing population of aged people living alone. A wearable and ambulatory monitoring device would be ideal meet the demand. A miniaturized finger ring sensor has been proved to be a promising candidate for cardiovascular system monitoring. In the previous stage, the miniaturized ring sensor has been proved to be feasible to be fabricated and provide the same function as the prototype developed in Phase I.

As is found in the first version prototype of the miniaturized sensor, the low frequency signal processing circuit and the high frequency transmission circuit cannot work very well on the same circuit board due to the heavy interference between them. Meanwhile, packaging the sensor into the ring configuration has not been taken into account in the first design. The distribution of the I/O connections on the board adds more difficulties for packaging. Besides, some of the design didn't consider the process of the fabrication, thus make the fabrication process more difficult and less effective.

Based on these considerations, a new design of the miniaturized ring sensor was developed during this half year, as shown in Figure 10. The purpose of this new version is to increase the reliability and robustness of the miniaturized ring sensor against the disturbances from the environment. Besides, it also makes the fabrication more efficient by introducing the systematic design methodology.

The new version has two circuit boards, which are signal processing and telemetry respectively. Four-layer PCB is adopted to reduce the size of the sensor. Two batteries are sandwiched between the two PCBs to supply the power to the two circuits. (It has been found that the two circuits have to be powered separately in order to work properly.) I/O connections are distributed on the edge of the boards, providing the connections for power supplies, LEDs and programming. Four screws will be used in the four ears on the board to provide mechanical fixtures for the boards. All the circuitry on the boards will be protected by optical epoxy after fabrication and debugging.

In this report, the detail design of the new version of the miniaturized ring sensor will be discussed. Issues arising from the design will be introduced and the solutions are provided.
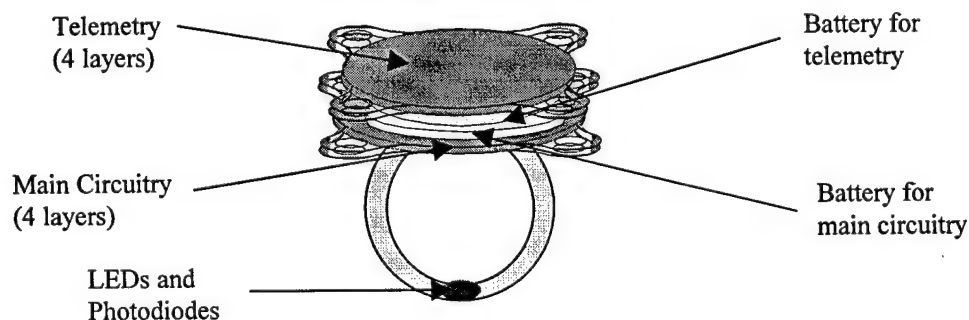

Figure 10: New design of miniaturized ring sensor

3.2 Issues on Design

The design of the miniaturized ring sensor has been improved by adding the following new features:

3.2.1 Packing all the components into a ring configuration

One of the most difficult issues in the design is how to pack all the components into a ring configuration. As we designed in the first version, all the circuitry are put into one piece of printed circuit board which is glued on the ring, where the LEDs and photo diodes sit. The disadvantage of the design is the difficulty to assemble the batteries. Since the two circuits (analog signal processing and telemetry) have to use separated power supply to work properly, two batteries have to be used to support the circuits. Meanwhile, the two circuits have to share common ground to get the correct signals during transmission. Another issue rising from the one-board design is that the high frequency (MHz) telemetry cannot work very well within the same board with the low frequency (KHz~10Hz) analog circuit in the current design.

The new design provided a solution to the problem.

- The two circuits, signal processing and telemetry, are separated into two PCBs to reduce interferes between them. As shown in Figure 11, the telemetry is on the top of the sensor while the main circuitry is on the bottom. They provide a supporting frame for the batteries and the ring.
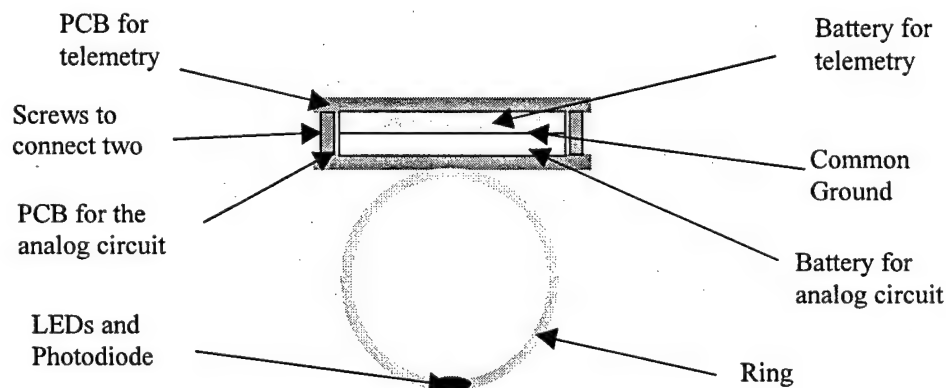

Figure 11: Side view of miniaturized ring sensor

- The batteries, therefore, are sandwiched between the two circuit boards, of which the sides have a large conducting area to connect with the battery, as shown in Figure 12.
- There are four ears on the board, which, by connection pins, provide the support to the boards. The schematic of the ear-board is shown in Figure 13.
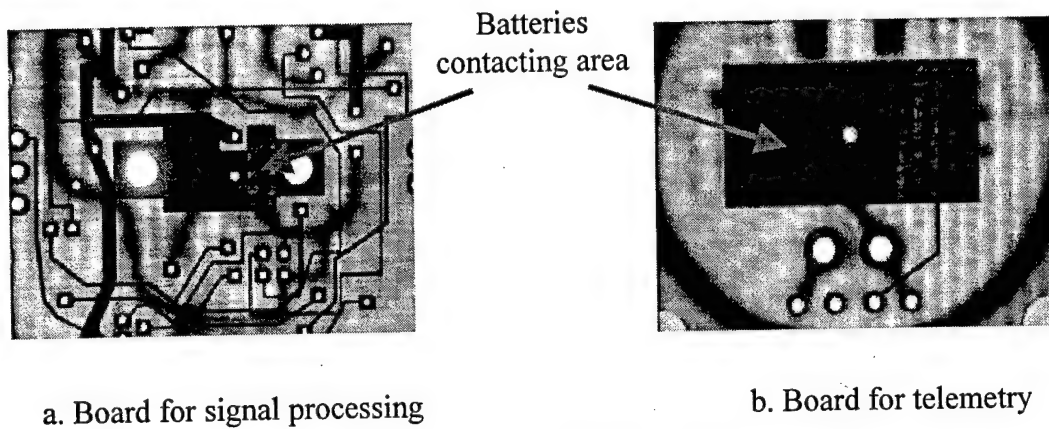
14

Batteries contacting area

a. Board for signal processing   b. Board for telemetry

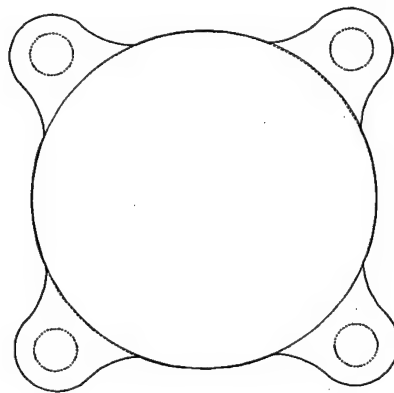**Figure 12: PCBs, the reverse side (battery connections)**

Figure 13: New design of the PCBs, ears providing mechanical supports

3.2.2 Taking full advantage of printed circuit board

The two options for fabricating the circuit board are printed circuit board (PCB) and ceramic substrate. As we have stated in the last progress report, ceramic substrate is not suitable for small volume prototype development. Thus we choose printed circuit board to meet the requirement of short lead-time and low cost.

However, we didn't take the full advantage of PCB in the first version. One of the most noticeable advantages of PCB is the usage of multiple layers to reduce the circuitry on one layer, therefore reduce the size of the ring sensor.

As we can see on the diagram of the previous version, more than half of the board are occupied by power supplies and grounds lines. These lines provided powers to the ICs, such as microprocessor, Op Amps, etc., which are widely distributed on the board. Multiple-layer design will significantly reduce the size of PCB.

In the current version of the pattern, we adopt the four-layer PCB as shown in Figure 14. The four layers are designed in the following way: the middle two layers are power supply and ground respectively. They are connected to the corresponding points in the

15

main circuitry to provide the power and common the ground. The top layer is the main circuitry, where all the parts are sitting. The bottom layer offers the connections to the battery.
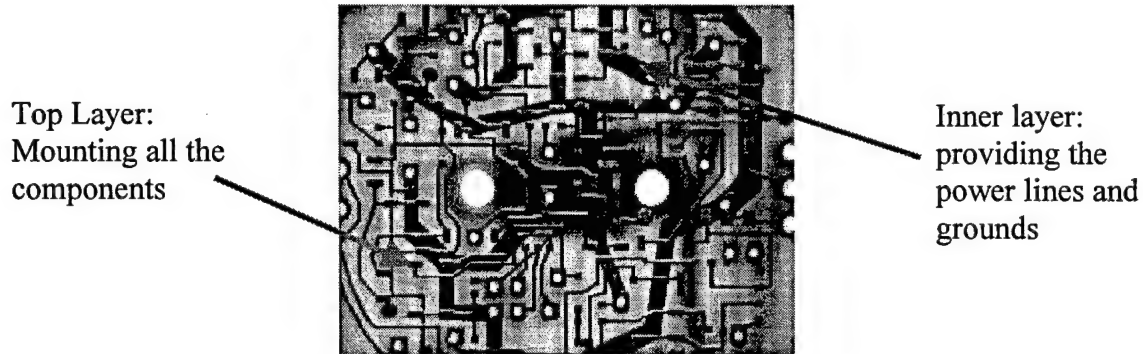
Top Layer:
Mounting all the
components

Inner layer:
providing the
power lines and
grounds

Figure 14: Multiple-layer printed circuit board (Main circuit board)

3.2.3 Optimizing the design for assembly and debugging

The concept of design for manufacture and assembly rises from the manufacturing industry to reduce the cost for the production. The principle idea is the integration of the whole process of design, manufacturing, assembly and testing for a product. Manufacturing and assembling of a product are considered during the early stage of product design so that a product is designed to not only meet functionality and specifications, but also can be manufactured and assembled economically and with relative ease.

The methodology of design for assembly (DFA) can be and have been adopted into the design of miniaturized ring sensor. The new design minimizes the complexity of the fabrication and assembly of the sensor; therefore the ring sensor can be fabricated by relative less effort and time. The application of the concept reduces the production cost for the ring sensor.

Meanwhile, the difficulties for the miniaturized ring sensor exist in the debugging and testing of the sensor. All the components are too tiny to be probed by even mini-grabbers (0.75mm diameter). Besides, all the ICs used in the sensor are in die form and bonded by 25µm golden wire, which is impractical to be probed. We imbedded the similar idea of DFA for this stage, which we call design for debugging (DFD). The idea is that we take the debugging into account during the design so that debugging and testing the mini-sensor is not a headache anymore.

All the revisions made according to the idea of design for assembly and debugging are discussed in the following section.

3.3 Issues on Assembly and Debugging

3.3.1 Pattern design for ICs

Assembly is one of the most critical tasks in the fabrication of the miniaturized ring sensor since all ICs in die form have to be wire-bonded. Making sure that all bonds are in right shape is the first and important step to assure successful assembly.

Most important of all, the golden wire should not touch the edge of a die when it crosses the ICs in order to avoid the potential of short-circuit on the die due to the conducting golden wire's touching the silicon. In the first design, based on the requirement of squeezing the size of the board, the patterns for the Op Amp were evenly distributed around the chip. As shown in Figure 15, the connections on pin 5, 6 and 7 have a long 'travel' on the chip, which increases the possibility of short-circuit.

In the new design, the pattern is re-arranged around the chip to minimize the distance between the connections, therefore reduce the travelling distance of the golden wire. Consequently, it reduces the chance of short-circuit and makes the wire bonding easier as well.
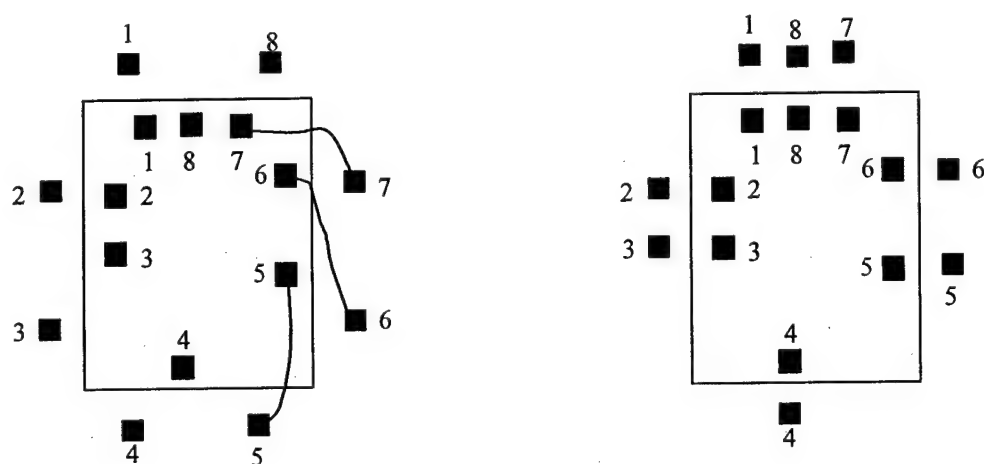

Figure 15: Redesign of the pattern for Op Amp MAX407

3.3.2 Connections for input and output of the sensor

There are totally 13 I/O connections on the main circuitry: five of them for CPU programming, five for the LEDs and photodiodes, three for power supplies. They have to be distributed along the edge of the board in order to make the ring package possible. The I/Os are currently designed as follow to make the packaging easier:
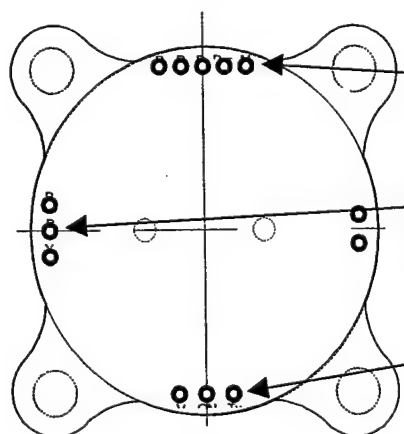
The five connections for programming are put together so that they can be connected to the corresponding pins of an EPROM writer.

The five connections to the LEDs and photodiodes are distributed on the two sided of the board since they come from the ring side in two directions.

The three connections for the power supply are put on the opposite side of the connections for programming. They

17

are connected by the two inner layers to supply the power to the circuit.

3.3.3 Circuit debugging

As we have stated above, it is difficult to debug and test the miniaturized ring sensor since there is no way to probe on the dies. Therefore in the new design, several testing pins have been added into the patterns, which make the debugging much easier.

We are mainly concerned about the output of the first stage Op Amp (to see if the photo diode grabs the correct signal and if the circuit is working); outputs of the switch (to see if the program is running correctly), outputs of the four Op Amps afterwards (to see if the signal processing is working properly). Since the first stage Op Amp we are using is surface mount type, it is relatively easy to probe. Therefore, we only design the debugging pads for those die-form ICs. The round solid pads on the top layer of the main circuitry are specially designed for debugging.

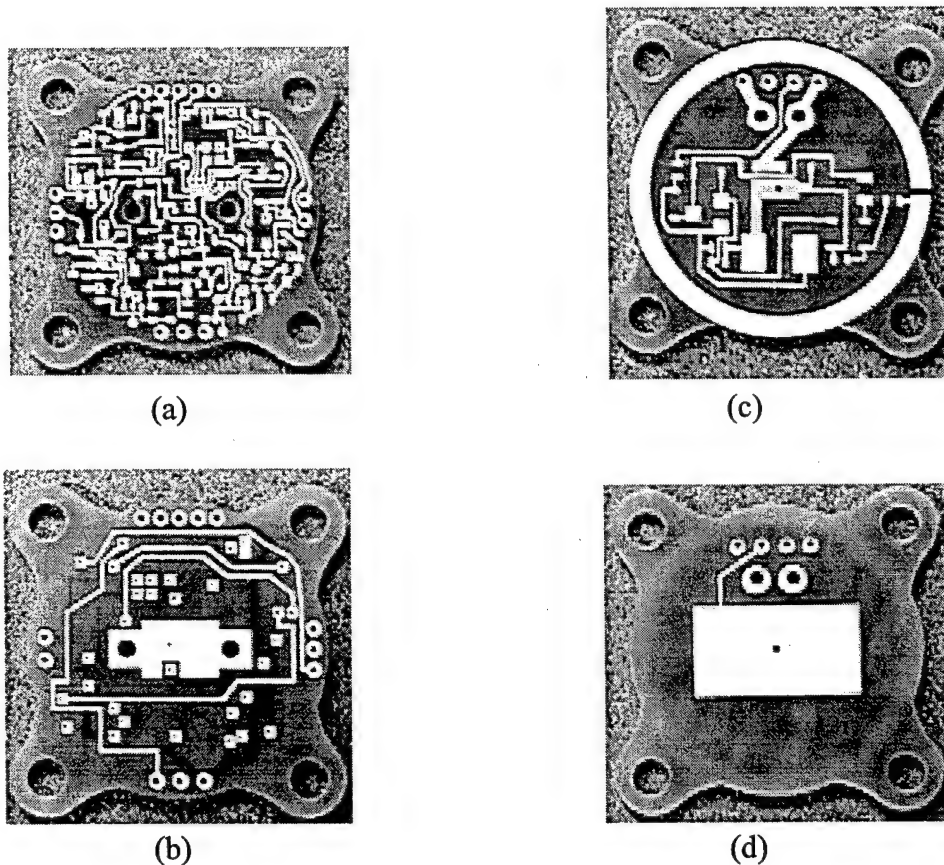The picture of the new designed circuitry of the miniaturized ring sensor is shown in Figure 16.

(a)

(c)

(b)

(d)

Figure 16: Pictures of the new circuit boards of the miniaturized ring sensor: (a)Main circuitry (Top View) (b) Main circuitry (Bottom View) (c) Telemetry (Top View) (d) Telemetry (Bottom View)

3.3.4 In-circuit Programming

The ring sensor is designed so that it can be re-programmed on the circuit board. This feature is extremely critical during the prototype development since frequent modifications of the program are needed.

A large EPROM eraser is needed to assure the program stored in the EPROM can be erased completely. The whole package of the sensor including the CPU has to be put into the eraser so that strength of the ultraviolet radiation will be large enough.

3.4 Conclusion and Future Work

A new design of the miniaturized ring sensor based on the concept of DFA and DFD was developed, which effectively reduces to the size of the sensor without sacrificing any of the functionality. Meanwhile, the hardware design was improved so that it eliminates many potentials of mistakes during the fabricating and testing of the ring sensor.

In the future, more functionality, including ambient light elimination and advanced signal processing algorithms, will be implemented in the miniaturized ring sensor.

4 Conclusions

A new algorithm for rejecting the noise caused by motion artifact and ambient light in ring sensor application was presented using the autocorrelation functions. The autocorrelation function is a random signal processing technique that can be used to reconstruct the original waveform in spite of the presence of severe noise with a bad SNR. A theoretical explanation of this method was presented with a numerical simulation, and the experimental data obtained by the ring sensor was processed with this technique. The experimental result shows that a combination of the classical low pass filtering technique and this autocorrelation method can be effectively used for noise reduction caused by motion artifact and ambient light although it has some limitations in its application. In the future, this signal correlation technique can be combined with input modulation technique and is expected to give an even better result in reconstructing the original waveform of the heart beat.

A new design of the miniaturized ring sensor for twenty-four-hour patient monitoring was also described in detail. The problems that arise from old version of the design were presented, and the solutions to these problems were provided. The methodology of design for assembly and debugging was also introduced and the improved design and fabrication of the miniaturized finger ring sensor were described in detail.
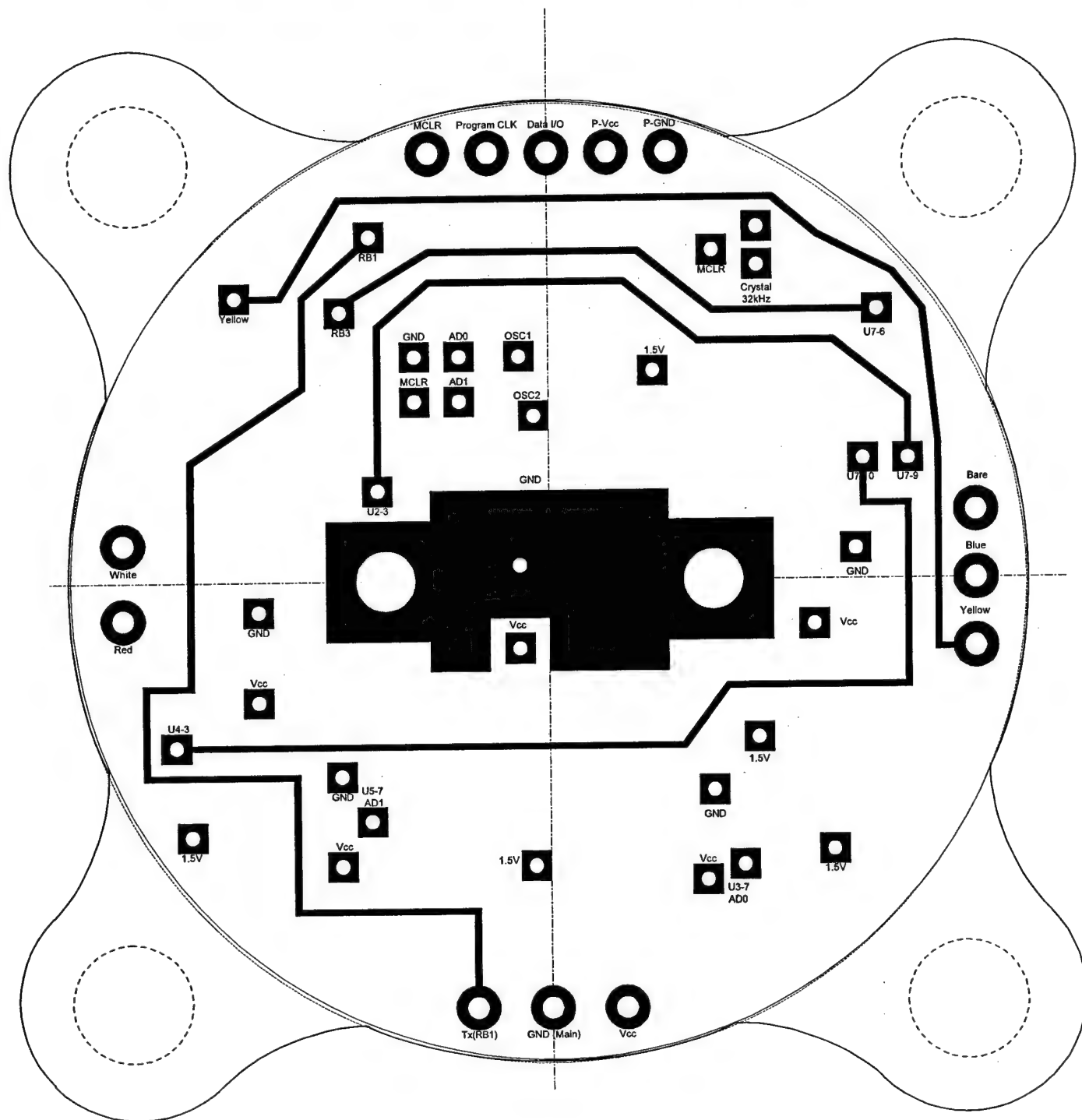
Appendix 1: Top Layer of the Main circuitry
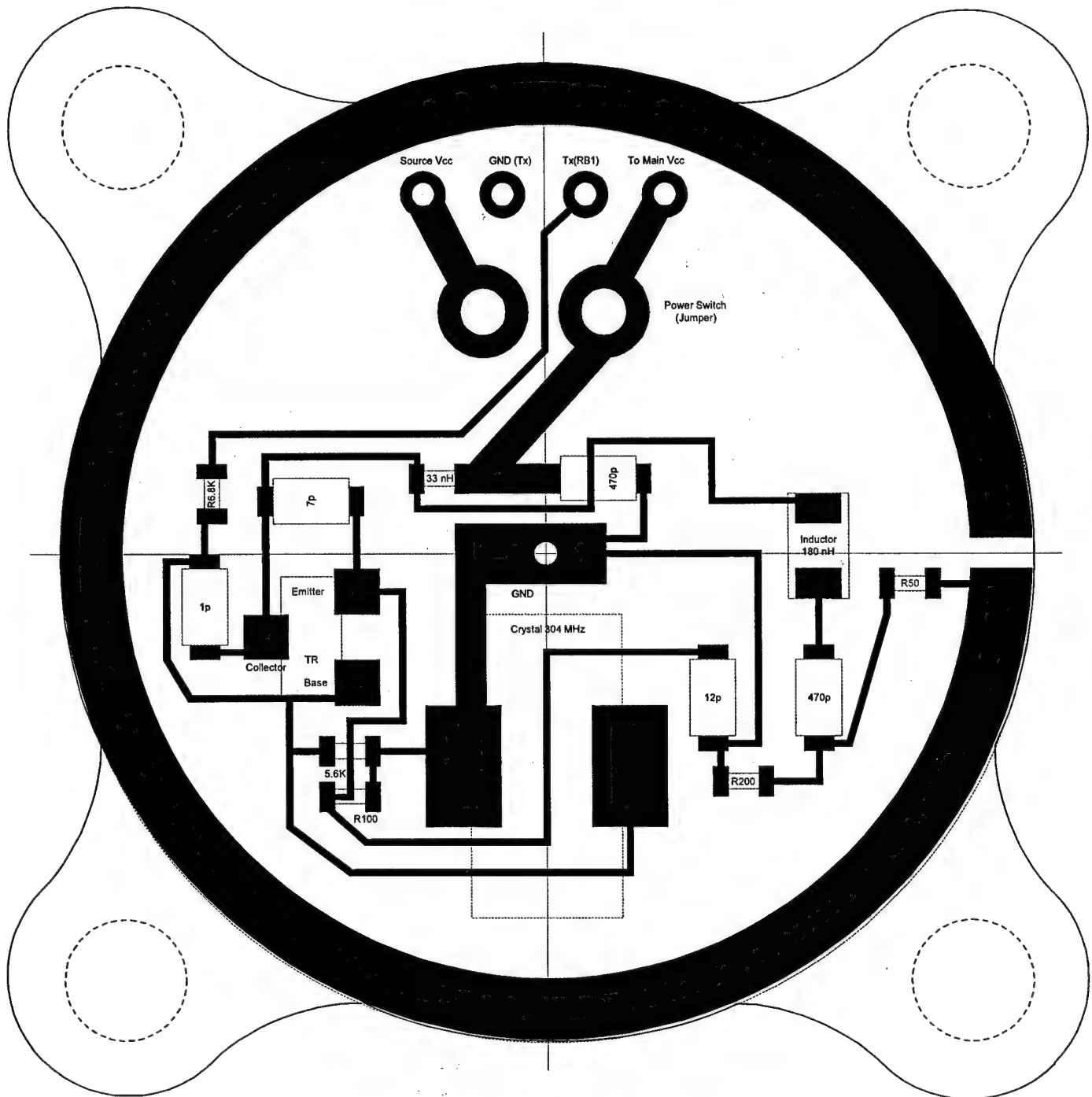
Main Layer 1 : General Circuitry
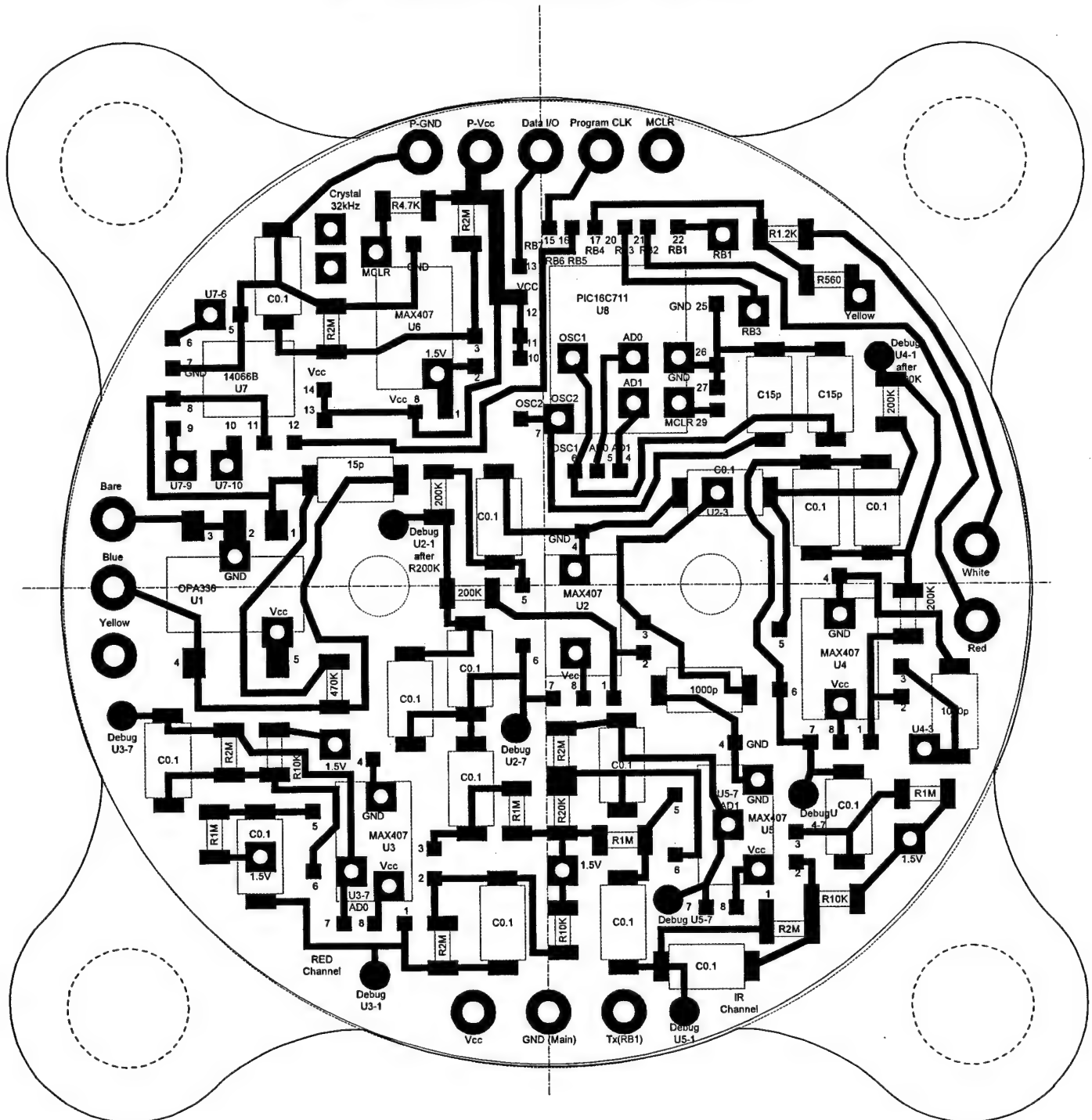
Appendix 2: Bottom Layer of the Main circuitry

Main Layer 4 (Rev) : LED Yellow Wire, Tx, U7-9, U7-10,

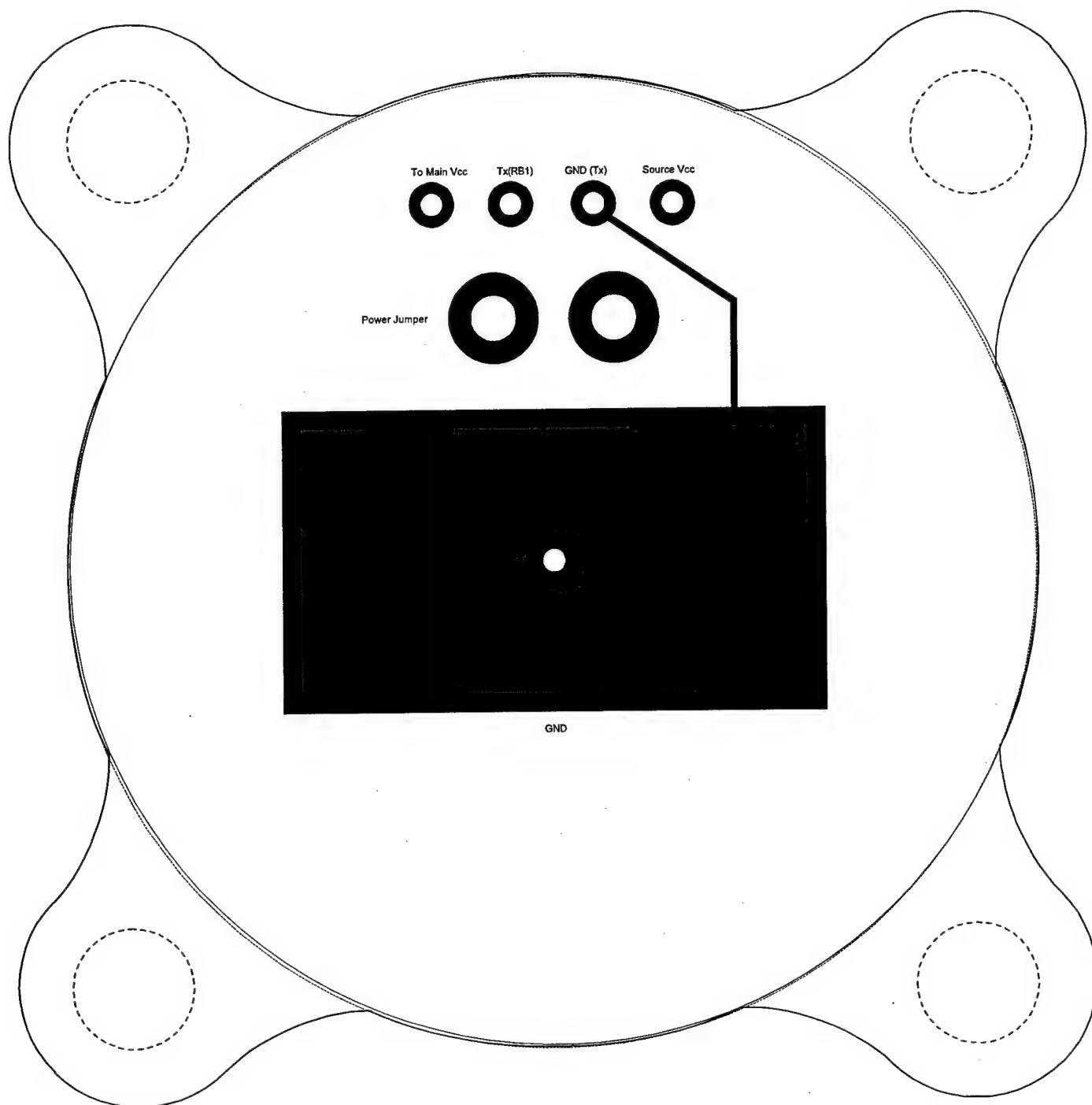# Appendix 3: Top Layer of Telemetry

Tx Layer 1 : General Tx Circuit

Appendix 4: Bottom Layer of Telemetry

Tx Layer 2 (Rev) : GND

To Main Vcc   Tx(RB1)   GND (Tx)   Source Vcc

Power Jumper

GND

Patient Monitoring and Diagnosis

CHAPTER 3

Sensor Fusion for Continuous Monitoring of Hemodynamic States
B-H Yang, H. Asada

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Sensor Fusion for Continuous Monitoring of Hemodynamic States

Boo-Ho Yang and Harry Asada

1. Introduction

As the population of aged people increases, technologies for home healthcare automation are badly needed. Considering that heart disease is a prevalent cause of death in the modern society, a technology for constant assessment of the patient's cardiovascular system at home is a key component for high-quality home healthcare. Particularly, for detecting and predicting a variety of cardiovascular disorders, precise assessment of hemodynamic variables such as the arterial blood pressure and flow pulses is necessary. However, there are no devices or sensors available to directly monitor the time-varying hemodynamic states continuously and noninvasively. Therefore, it is quite important to develop an indirect, model-based approach that allows the hemodynamic state to be estimated based on available sensor information.

In this project, we proposed a sensor fusion approach to continuous monitoring of the hemodynamic state. The idea of sensor fusion is to process and merge information gathered by multiple sources and sensors to provide a better insight into the phenomena under consideration. It is expected that by integrating sensor signals from multiple available sources with a hemodynamic model, we will be able to estimate the important hemodynamic variables that cannot be measured directly. To prove the above argument, the theory underpinning sensor fusion for the hemodynamic system is derived by formulating a linear state observer problem.

The objective of this report is to provide a preliminary work on the sensor fusion approach to continuous monitoring of the hemodynamic state. First, a two-dimensional mathematical model of an arterial blood flow is derived. Then, the model is linearized into a state-space equation for analysis of the dynamic behavior. The observability of the system for a variety of combinations of available sensors is examined.

The sensor fusion approach is applied to a ring sensor. The ring sensor is a compact, wearable monitoring system in a ring configuration that can be comfortably worn by the patient twenty-four hours a day. The ring is equipped with LEDs and photodetectors for monitoring the volumetric changes of the arterial blood flow in a finger base. It is expected that the above analysis of the observability would give us a guideline for designing a new ring sensor system.
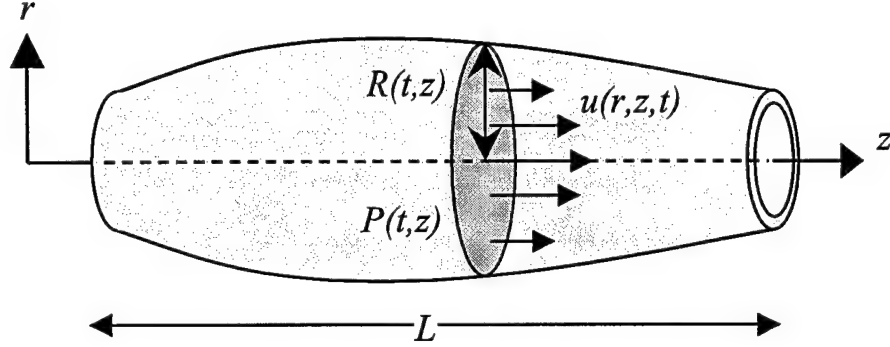
1

Figure 1: A segment of a viscoelastic artery with length of L

2. Hemodynamic Modeling of Arterial Blood Flow

Mathematical Model of Arterial Flow

Many hemodynamic models have been developed for the study of pressure wave propagation along an artery. In this report, we apply a mathematical framework developed by Belardinelli and Cavalcanti [1], which describes a two-dimensional nonlinear flow of Newtonian viscous fluid moving in a deformable tapered tube.

A small segment (distance of L) of a small artery such as a digital artery is considered in this report as shown in Figure 1. We assume that the arterial vessel is rectilinear, deformable, thick shell of isotropic, incompressible material with a circular section and without longitudinal movements. Blood is an incompressible Newtonian fluid and flow is axially symmetric. Two-dimensional Navier-Stoke equations and continuity equation for a Newtonian and incompressible fluid in cylindrical coordinate (r, θ, z) are:

$$\frac{\partial u}{\partial t} + w\frac{\partial u}{\partial r} + u\frac{\partial u}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial z} + \nu\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2}\right) \tag{1}$$

$$\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial r} + u\frac{\partial w}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial r} + \nu\left(\frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r} + \frac{\partial^2 w}{\partial z^2} - \frac{w}{r^2}\right) \tag{2}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(rw) + \frac{\partial u}{\partial z} = 0 \tag{3}$$

where P denotes pressure, ρ density, ν kinematic viscosity, and $u=u(r,z,t)$ and $w=w(r,z,t)$ denote the components of velocity in axial (z) and radial (r) directions respectively, as shown in the above figure. Let $R(z,t)$ denote the inner radius of the vessel and define a new variable:

$$\eta = \frac{r}{R(z,t)} \tag{4}$$

2

We also assume that the pressure P is uniform in within the cross section so that P is independent of the radial coordinate, η, i.e. $P=P(z,t)$. We can rewrite the above equations in a new coordinate (η, θ, z) as

$$\frac{\partial u}{\partial t} + \frac{1}{R}(\eta(u\frac{\partial R}{\partial z} + \frac{\partial R}{\partial t}) - w)\frac{\partial u}{\partial \eta} + u\frac{\partial u}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial z} + \frac{\nu}{R^2}(\frac{\partial^2 u}{\partial \eta^2} + \frac{1}{\eta}\frac{\partial u}{\partial \eta}) \tag{5}$$

$$\frac{\partial w}{\partial t} + \frac{1}{R}(\eta(u\frac{\partial R}{\partial z} + \frac{\partial R}{\partial t}) - w)\frac{\partial w}{\partial \eta} + u\frac{\partial w}{\partial z} = \frac{\nu}{R^2}(\frac{\partial^2 w}{\partial \eta^2} + \frac{1}{\eta}\frac{\partial w}{\partial \eta} - \frac{w}{\eta^2}) \tag{6}$$

$$\frac{1}{R}\frac{\partial w}{\partial \eta} + \frac{w}{\eta R} + \frac{\partial u}{\partial z} - \frac{\eta}{R}\frac{\partial R}{\partial z}\frac{\partial u}{\partial \eta} = 0 \tag{7}$$

where we assume:

$$\frac{\partial^2 u}{\partial z^2} \ll 1, \quad \frac{\partial^2 w}{\partial z^2} \ll 1, \quad \frac{\partial P}{\partial r} \ll 1$$

The boundary conditions for the above equations in η axis are:

$$w(\eta,z,t)\big|_{\eta=0} = 0, \quad w(\eta,z,t)\big|_{\eta=1} = \frac{\partial R}{\partial t}, \quad u(\eta,z,t)\big|_{\eta=1} = 0, \quad \frac{\partial u}{\partial \eta}\big|_{\eta=0} = 0 \tag{8}$$

The basic idea of this hemodynamic modeling provided by [1] is to assume that the velocity profile in the axial direction can be expressed as the following polynomial form:

$$u(\eta,z,t) = \sum_{k=1}^{N} q_k(\eta^{2k} - 1) \tag{9}$$

The velocity profile in the radial direction is also expressed as:

$$w(\eta,z,t) = \frac{\partial R}{\partial z}\eta w + \frac{\partial R}{\partial t}\eta - \frac{\partial R}{\partial t}\frac{1}{N}\eta\sum_{k=1}^{N}\frac{1}{k}(\eta^{2k} - 1) \tag{10}$$

For simplicity, we choose $N=1$ in this report such as

$$u(\eta,z,t) = q(z,t)(\eta^2 - 1) \tag{11}$$

$$w(\eta,z,t) = \frac{\partial R}{\partial z}\eta w + \frac{\partial R}{\partial t}\eta - \frac{\partial R}{\partial t}\eta(\eta^2 - 1) \tag{12}$$

By plugging eqs.(11) and (12) into eqs.(5) and (7), we obtain the dynamic equations of $q(z,t)$ and $R(z,t)$ as:

$$\frac{\partial q}{\partial t} - \frac{4q}{R}\frac{\partial R}{\partial t} - \frac{2q^2}{R}\frac{\partial R}{\partial z} + \frac{4\nu}{R^2}q + \frac{1}{\rho}\frac{\partial P}{\partial z} = 0 \tag{13}$$

3

$$2R\frac{\partial R}{\partial t}+\frac{R^2}{2}\frac{\partial q}{\partial z}+q\frac{\partial R}{\partial z}=0 \qquad (14)$$

Complete derivations of the above equations are found in [1]. Let us define cross-sectional area $S(z,t)$ and blood flow $Q(z,t)$ as

$$S=\pi R^2, \quad Q=\iint_S u\,d\eta =\frac{1}{2}\pi q R^2$$

Then, eqs.(13) and (14) cab be re-written in terms of Q and S as:

$$\frac{\partial Q}{\partial t}-\frac{3Q}{S}\frac{\partial S}{\partial t}-\frac{2Q^2}{S^2}\frac{\partial S}{\partial z}+\frac{4\pi\nu}{S}Q+\frac{S}{2\rho}\frac{\partial P}{\partial z}=0 \qquad (15)$$

$$\frac{\partial S}{\partial t}+\frac{\partial Q}{\partial z}=0 \qquad (16)$$

Viscoelastic Model of Arterial Wall

To study the hemodynamics of arterial blood flow, a modeling of the viscoelastic behavior of the arterial wall is essential. In this report, we will derive a constitutive law of the arterial wall from stress-strain relationship of the material. Let σ_θ and σ_t be the circumferential stress and tangential stress respectively as shown in Figure 2. Ignoring the inertia of the arterial wall and the external pressure, equilibrium with the blood pressure gives:

$$PR=\sigma_\theta e-\sigma_t eR\frac{\partial^2 R}{\partial z^2} \qquad (17)$$

where $R(z,t)$ and e are the radius of the arterial vessel and the thickness of the arterial wall respectively.
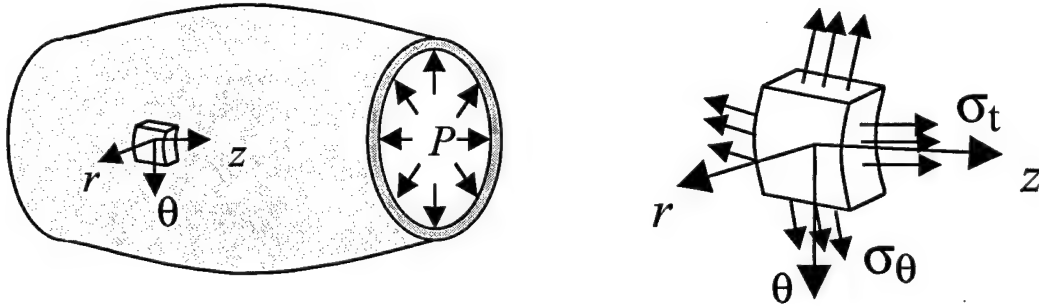

Figure 2: State of stress in a thin-walled, elastic blood vessel

From the geometric compatibility of the blood vessel, we get an expression of strains such as

4

$$\varepsilon_\theta = \frac{R - R_0}{R_0}, \qquad \varepsilon_t = \sqrt{1 + \left(\frac{\partial R}{\partial z}\right)^2} - 1 \tag{18}$$

where ε_θ and ε_t are circumferential and tangential strains respectively and a constant R_0 is the radius of the artery when $P(z,t)=0$ and the system is in a steady state.

The most widely used model to describe the viscoelastic properties of the arterial wall is the Kelvin-Voigt model, in which the stress-strain relationship is described as:

$$\sigma_\theta = E\left(\varepsilon_\theta + \eta \frac{\partial \varepsilon_\theta}{\partial t}\right), \quad \sigma_t = E\left(\varepsilon_t + \eta \frac{\partial \varepsilon_t}{\partial t}\right) \tag{19}$$

in which E is the elastic modulus and η is the damping coefficient. By plugging eqs.(18) and (19) with $S_0 = \pi R_0^2$ and eliminating second and higher order terms, we get the following equation as the viscoelastic constitutive law of the arterial wall:

$$P = \frac{\sqrt{\pi} E e}{2 S_0 \sqrt{S}} \left(S + \eta \frac{\partial S}{\partial t} - S_0\right) \tag{20}$$

Development of a State-State Model

The above nonlinear, partial differential equations given in eqs.(15), (16) and (20) are discretized and transformed into a linear state equation using a finite-difference method. First, the segment of the artery (length L) is equally divided by N grids with a step size of $\Delta z = L/(N-1)$. The mesh points in the finite difference grids are represented by j where $j=1,2,\cdots,N$. The derivatives in terms of z are substituted by differences such as:

$$\frac{\partial S_i}{\partial z} = \frac{S_{i+1} - S_i}{\Delta z}, \quad \frac{\partial P_i}{\partial z} = \frac{P_{i+1} - P_i}{\Delta z}, \quad \frac{\partial Q_i}{\partial z} = \frac{Q_i - Q_{i-1}}{\Delta z} \tag{21}$$

The constitutive law given in eq.(20) is modeled such that the viscoelasticity applies only at mesh points. An example of the discretization when $N=4$ is shown in Figure 3. Using the above equations, the hemodynamic model given in eqs.(15) and (16) can be discretized as

$$\frac{dQ_i}{dt} + \frac{3Q_i}{S_i}\frac{Q_i - Q_{i-1}}{\Delta z} - \frac{2Q_i^2}{S_i^2}\frac{S_{i+1} - S_i}{\Delta z} + \frac{4\pi\nu}{S_i}Q_i + \frac{S_i}{2\rho}\frac{P_{i+1} - P_i}{\Delta z} = 0 \tag{22}$$

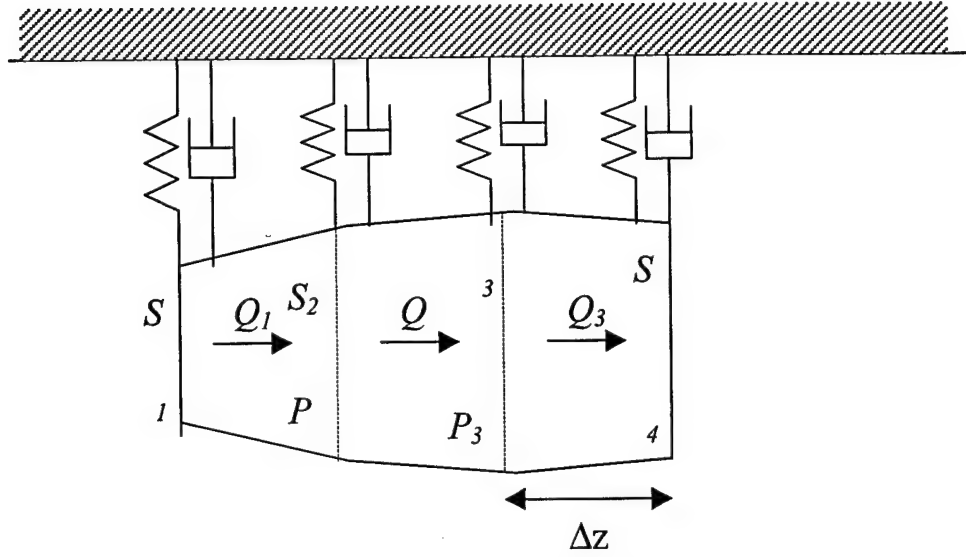$$\frac{dS_i}{dt} = \frac{Q_i - Q_{i-1}}{\Delta z} \tag{23}$$

5

Figure 3: Discretization of the hemodynamic model

We define the state variables of the system as

$$x = (Q_1, \cdots, Q_{N-1}, S_1, \cdots, S_N)^T : (2N-1) \times 1 \tag{24}$$

From the continuity equation given by eq.(16) and the constitutive law of the arterial wall given by eq.(20), the pressures Pi can be expressed in terms of the above state variables as:

$$P_i = \frac{\sqrt{\pi} Ee}{2S_0\sqrt{S_i}}(S_i + \eta \frac{dS_i}{dt} - S_0) = \frac{\sqrt{\pi} Ee}{2S_0\sqrt{S_i}}(S_i - \frac{\eta}{\Delta z}(Q_i - Q_{i-1}) - S_0) \quad for \; i = 1, 2, \cdots, N \tag{25}$$

To complete the discretization of the above hemodynamic model, the boundary conditions at both ends of the segment must be defined appropriately. From the viewpoint of the dynamic modeling, the boundary conditions determine the sources of the motion within the defined segment of the artery. In this report, we consider the blood flows at both ends (Q_0 and Q_N) as the sources of the motion and define $u(t)$ as:

$$u = (Q_0, Q_N)^T : 2 \times 1 \tag{26}$$

For further analysis of the nature of the hemodynamic behavior of the arterial flow, we linearized the above equations and derived a linear state equation for the state variables given in eq.(24) and the system inputs given in eq.(26) as follows:

$$\frac{dQ_i}{dt} + \frac{4\pi\nu}{S_0}Q_i - \frac{\sqrt{\pi} Ee\eta}{4\rho\Delta z^2 \sqrt{S_0}}(Q_{i+1} - 2Q_i + Q_{i-1}) + \frac{\sqrt{\pi} Ee}{4\rho\Delta z\sqrt{S_0}}(S_{i+1} - S_i) = 0 \; for \; i = 1, \cdots, N-1 \tag{27}$$

$$\frac{dS_i}{dt} = \frac{Q_i - Q_{i-1}}{\Delta z} \; for \; i = 1, 2, \cdots, N \tag{28}$$

6

In a matrix form, the above state equations can be expressed as:

$$\dot{x} = Ax + Bu \qquad (29)$$

where

$$A = \begin{bmatrix} -\dfrac{4\pi v}{S_0} I_{N-1} - \dfrac{\sqrt{\pi}\, E e \eta}{4\rho \Delta z^2 \sqrt{S_0}} J_{N-1} & \dfrac{\sqrt{\pi}\, E e}{4\rho \Delta z \sqrt{S_0}} H_{N-1} \\ -\dfrac{1}{\Delta z} H_{N-1}{}^T & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} \dfrac{\sqrt{\pi}\, E e \eta}{4\rho \Delta z^2 \sqrt{S_0}} B_1 \\ \dfrac{1}{\Delta z} B_2 \end{bmatrix}$$

$$J_{N-1} = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & & & \vdots \\ 0 & -1 & 2 & -1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -1 & 2 & -1 & 0 \\ \vdots & & & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & 0 & -1 & 2 \end{bmatrix} \quad : (N-1) \times (N-1)$$

$$H_{N-1} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & & & \vdots & 0 \\ 0 & 0 & 1 & -1 & \ddots & & \vdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \\ \vdots & & \ddots & 0 & 1 & -1 & 0 & 0 \\ \vdots & & & 0 & 0 & 1 & -1 & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \quad : (N-1) \times N$$

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \quad : (N-1) \times 2, \quad B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & -1 \end{bmatrix} \quad : N \times 2$$

3. State Observer and Sensor Fusion

A state observer is a dynamic system whose state variables are the estimates of the state variables of another dynamic system. Observers are popularly used to estimate unknown state variables that cannot be measured directly with limited measurements of the system.

In this project, as a sensor fusion approach, we develop a state observer to estimate time-varying internal variables such as blood pressures.

To formulate a state observer for the above hemodynamic system given by eq.(29), the observation equation must be defined based on the instrumentation technology to be used. In the ring sensor project, we have built a new concept of the ring sensor, called "hyper ring." The new sensor system has two rings along a finger base and each ring is equipped with LEDs and photodetectors for photoplethysmography. The two photoplethysmograms can be simply defined as cross-sectional areas at both ends of the arterial segments as:

$$y_1(t) = S_1(t), \quad y_2(t) = S_N(t)$$

Four electrodes are also installed in the two-ring configuration for electrical impedance plethysmography (EIP). The EIP is known to provide absolute measurement of volumetric change of an arterial segment. Therefore, the output of the EIP in terms of the state variables can be described as:

$$y_3(t) = V(t) = (S_1 + S_2 + \cdots + S_{N-1})\Delta z$$

Defining $y(t) = [y_1(t), y_2(t), y_3(t)]^T$, the observation equation can finally be defined as

$$y(t) = Cx(t) \tag{30}$$

where

$$C = \begin{bmatrix} \overbrace{0 \quad \cdots \quad 0}^{N-1} & \overbrace{1 \quad 0 \quad \cdots \quad 0 \quad 0}^{N} \\ 0 \quad \cdots \quad 0 & 0 \quad 0 \quad \cdots \quad 0 \quad 1 \\ 0 \quad \cdots \quad 0 & \Delta z \quad \Delta z \quad \cdots \quad \Delta z \quad 0 \end{bmatrix}$$

To construct an state observer for the system given by eqs.(29) and (30), we apply Luenberger's method. First, we define $\hat{x}(t)$ to be an estimate of the state variables $x(t)$. Assuming the input $u(t)$ is always available, the observer is described as

$$\dot{\hat{x}} = A\hat{x} + Bu + K(y - \hat{y}) \tag{31}$$
$$\hat{y} = C\hat{x} \tag{32}$$

where $\hat{y}(t)$ is the estimated measurement and K is the error feedback gain matrix of the observer. For an appropriate choice of K, it is known that the above estimate of the states converges to the true state variables if the system above is observable. To prove the convergence under the assumption that the system is observable, let us define the state estimate error as

$$\tilde{x} = x - \hat{x} \tag{33}$$

8

Then, from eqs. (29) and (31), we get

$$\dot{\tilde{x}} = (A - KC)\tilde{x} \tag{34}$$

Therefore, by choosing K so that the real parts of all the eigenvalues of (A-KC) become negative, the estimate error $\tilde{x}(t)$ converges to zero. Namely, the state variables that cannot be measured can be precisely estimated from output measurements.

The main issue in the above observer problem is whether the system given in eqs.(29) and (30) is observable or not. If the system is not observable, we cannot construct an observer to estimate the whole state variables. As it is found in the next section, the above system is not observable based on the available measurements given in eq.(30). However, the observability analysis to be provided in the next section will give us a useful guideline about how to modify the ring sensor system to make the system observable.

4. Observability Analysis

There are many criteria for testing the observability of a system (e.g. [2] for a textbook). The most noticeable test for the observability is so called, "Algebraic Controllability Theorem," and it simply states:

A system (A, C) of order n is observable if and only if the rank of the observability test matrix

$$O = [C^T, A^T C^T, \cdots, (A^T)^{n-1} C^T] \tag{35}$$

is equal to n.

This is arguably the easiest criterion to test the observability of a system. However, this criterion fails to provide a useful insight about unobservable systems.

For more detailed analysis of the unobservable system, a test using eigenvectors of the system is extremely useful. The test called "PBH Test" states:

A system (A, C) of order n is unobservable if and only if there exists a vector $v \neq 0$ such that

$$Av = \lambda v, \quad Cv = 0 \tag{36}$$

In other words, the system is observable if and only if there is no eigenvector of A that is orthogonal to the observation matrix C.

If a system is unobservable, there are certain modes of the system which cannot be observed from the system's output (observer functions). The above PBH test simply provides which modes of the system are unobservable. The analysis of the unobservable modes would give us a useful guideline for a design of the sensor fusion system.

9

The above observer tests were applied to the hemodynamic model and the sensor system given in eqs.(29) and (30). It was immediately found from the these tests that the system is observable if N, the number of the grids for discretization, is 3, and is not observable if N is larger than 3. Howeveer, if N is too small, the discretization error would be significant and the approximated hemodynamic model would lose accuracy and fidelity in a great degree. Therefore, even for a small segment of the artery monitored by a ring sensor, we believe that we have to model the arterial segment with N equal to or higher than 4, and, thus, the hemodynamic system becomes unobservable with the given set of the sensor signals of the ring sensor. The main issue is how to make the system observable with additional measurements. To address the issue, we have to examine the unobservable behavior of the system dynamics.

To illustrate the unobservable modes of the system, let us show an example when $N=4$. The observability matrix O as given by eq.(35) when $N=4$ is singular and the dimension of the null space is one, which means the only one mode of the system is unobservable with the measurements of the ring sensor system. The eigenvector that fails the PBH test given by eq.(36) can be easily found as:

$$v_1 = \begin{bmatrix} 0 \\ \overline{Q} \\ 0 \\ 0 \\ -\overline{S} \\ \overline{S} \\ 0 \end{bmatrix} \tag{37}$$

where the constants $\overline{Q}$ and $\overline{S}$ are determined from values of the system parameters. The motion of the mode corresponding to the above eigenvector is illustrated in Figure 4.
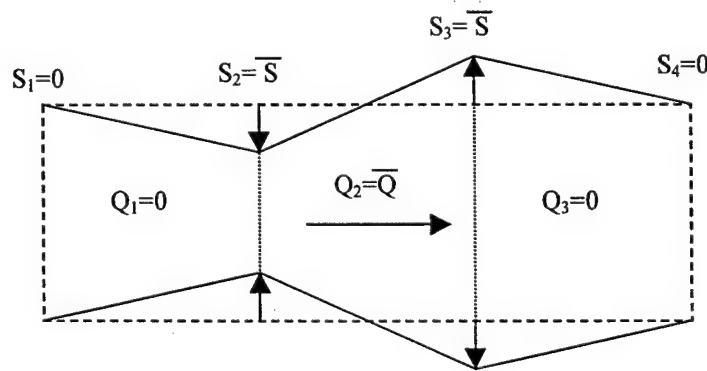

Figure 4: Unobservable mode

We can intuitively confirm that this particular mode cannot be observed from the observation functions given by eq.(30). Namely, it is obvious that output functions $y_1=S_1$, $y_2=S_4$, and $y_3=(S_1+S_2+S_3)\Delta z$ give zeros all the time while the system is excited only at this mode.

In other words, we can expect that the system could be observable if we add one more output that captures a motion of this mode, such as S_2, S_3, or Q_2. This expectation can be easily confirmed to be the case by formulating a new observation equation and test observability using one of the criteria. In fact, we do not need y_3, the impedance plethysmogram, if we have one of those additional measurements. Namely, if we have another photoplethysmopram or a flow measurement along the finger base, the hemodynamic system becomes observable and all the internal variables including blood pressures can be estimated. This would give us a useful guideline for designing a new type of the ring sensor for continuous monitoring of blood pressure.

5. Conclusions

In this reports, we proposed a new approach to monitoring hemodynamic states continuously and noninvasively using a sensor fusion technology. Sensor fusion allows internal state variables to be estimated by integrating available sensor information. An preliminary work on the underpinning theory for the sensor fusion was presented by formulating a state observer problem. The hemodynamic model of the arterial blood flow was derived and a linear state observer was developed based on the model. The theory of the sensor fusion was applied to a ring sensor that monitors volumetric changes of a digital artery at a multiple locations along the artery. The observability analysis of the ring sensor system provided a useful guideline for designing a new ring sensor for continuous monitoring of unmeasurable variables such as blood pressure.

References

1. E. Belardinelli and S. Cavalcanti, "A New Nonlinear Two-Dimensional Model of Blood Motion in Tapered and Elastic Vessels," Comput. Biol. Med., vol. 21, no. ½, pp. 1-13, 1991
2. T. Kailath, Linear Systems, Prentice-Hall, NJ, 1980

Patient Monitoring and Diagnosis

CHAPTER 4

SIMSUIT Projects

L. Jones, J. Tangorra, L. Sambol, E. Liu

Total Home Automation and Health Care Consortium

September 30, 1998

SIMSUIT Project

Lynette Jones

Principal Research Scientist

James Tangorra, Lisa Sambol

Graduate Research Assistants

Eric Liu

Undergraduate Research Assistant

Abstract: The SIMSUIT project is focused on the development of wearable health monitoring units that can measure different physiological variables such as heart rate, blood pressure, respiration rate, and core body temperature and evaluate the status of different sensory systems by perturbing them and measuring the responses to these perturbations. These monitoring systems must be lightweight, wireless, non-invasive and non-intrusive and so considerable effort is being devoted to miniaturizing the component elements and developing appropriate testing protocols. In this report research on the development of the vestibular-ocular testing apparatus will be described, together with a description of the use of this system to measure alertness and drowsiness. This apparatus is being developed for use in at least two different environments: as a clinical evaluation tool to examine the functioning of the human vestibular-ocular system, and as a device to measure the level of alertness/ drowsiness in human operators controlling vehicles or machines. An infrared tympanic thermometer is also being developed as part of the SIMSUIT project and this will be incorporated into the headphone-based vestibular-ocular testing device.

Vestibular-ocular Testing Device

The initial working configuration of our portable vestibular ocular reflex (VOR) testing device was completed in the spring of 1998. An inertial head perturber delivers small torque perturbations to a test subject's head while the test subject tracks a moving target with natural head and eye motions. An electro-oculograph (EOG), built from an in-house designed signal-conditioning board, tracks binocular eye motions, and magneto hydrodynamic rotational velocity sensors record head motions. Data are collected on the torque perturbations applied to the head, the rotational head velocity, eye position, and target position. Analysis is conducted with system identification software developed in Mathcad 8.0.
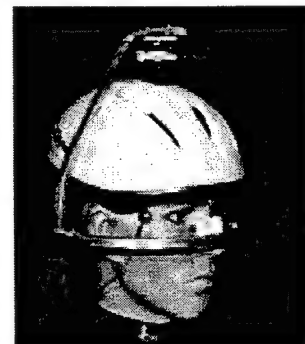
The performance of our testing equipment was first evaluated in a series of experiments designed to identify the impulse response function of the human head and neck. A parametric model was created, and values for mass, stiffness, and damping were calculated. Our calculations for head mass were similar to those found in published cadaver studies. Unfortunately, there do not appear to be any published data on relaxed neck properties, so we were unable to compare our findings with published results. As expected, the results confirmed the appropriateness of modeling the head and neck as a second-order damped mass spring. Though not especially rigorous, the tests showed our equipment and system identification protocols to be promising.

More rigorous testing of individual system components was necessary before the equipment could be used for vestibular testing or for the development of accurate data acquisition and analysis algorithms. Following the experiments on the head and neck, a calibrated damped rotational mass-spring apparatus was built that approximated the passive response of the human head and neck (Fig. 2). This allowed us to verify the accuracy of an impulse response estimated from stochastic system
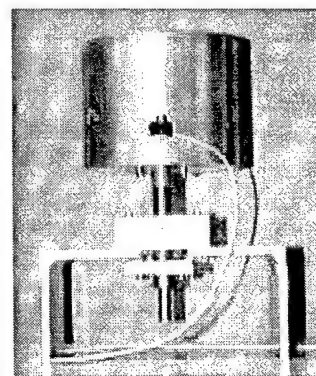
Fig 2: Damped mass-spring

identification to an impulse response determined from traditional input-output analysis.

A solid aluminum cylinder machined to a radius of 150 mm and a height of 110 mm models the head. This results in a rotational inertia of 0.0147 kg-m^2 about the vertical axis. In a study of 21 adult cadaver heads, Beier, Schuller, and Schuck found the rotational moment of inertia about the vertical axis to range from 0.0110 to 0.0198 kg-m^2, so our choice falls well within the bounds of "average" heads. The cylinder is mounted on a stainless steel shaft, the neck, which is supported by a shaft bearing set in a mounting stand. The end of the shaft is connected to a torque spring that provides damping and stiffness, and limits rotational movement to approximately +/- 30 degrees. The spring manufacturer claims a spring constant of 1.1 N-m, but the stiffness was statically measured to be 0.85 N-m. The spring was chosen to have a stiffness similar to the neck stiffness values measured in our previous testing of head and neck stiffness. The added mass of the neck shaft, mounting plate, torque spring, and rotational velocity sensor needed for testing, brought the rotational inertia of the system to 0.015 kg-m^2.

The impulse response of the mass spring system was evaluated by applying a pseudo-random binary (PRB) input torque to the rotational mass for 20 seconds. The PRB sequence was executed at frequencies from 10-30 Hz, which differed from trial to trial. A rotational velocity sensor attached to the rotational mass measured the resulting rotational motion. The impulse response (Fig. 3) was estimated by deconvolving the input torque's auto-covariance and the cross covariance of the input torque and output motion. A second-order parametric model was fitted to the impulse response using a Levenburg-Marquart minimization technique. The second-order parametric model did not fit the estimated impulse response as well as we expected. Individual parameters for rotational mass, damping, and stiffness fell within 20% of expected values, but the variance accounted for (VAF) between the actual output and the estimated output of the parametric model was approximately 90%, whereas we expected the VAF to be above 95%. It is suspected that the bandwidth of the input torque produced by the head actuator is a major cause of the inaccuracy. An accurate stochastic assessment of an impulse response requires that the frequency spectrum of the input be well distributed across the entire performance bandwidth of the system being tested. The helmet perturber's torque power spectrum is not well distributed; it is limited to 2-6 Hz (Fig. 4). This means that the rotational

system is only being stimulated from 2-6 Hz, and so the only information measured from the system is in this band. A 2-6 Hz band does not provide a good estimation of an impulse input to a system whose natural frequency is around 1 Hz. The limited bandwidth of the existing helmet perturber will be inadequate for accurate testing of the vestibular system which operates over a frequency range of less than 1 Hz to beyond 15 Hz. Low frequency testing is especially important, so the perturber must be modified to meet these needs. Though much simpler than the testing protocol and analysis required for the multi-input vestibular system, analyzing the rotational mass system provided important information about the performance of the helmet actuator, as well as experience with developing system identification algorithms. Most importantly, it was recognized that a new head perturber would have to be developed with the ability to test accurately the vestibular system across its full bandwidth.
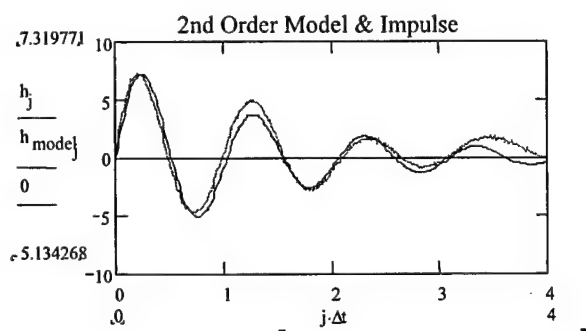
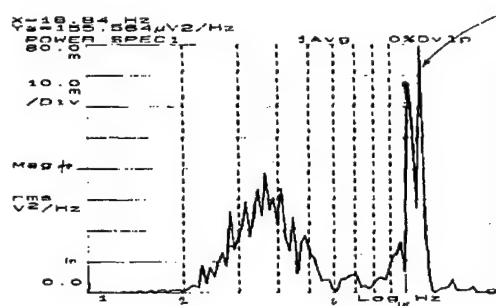Fig 3: Impulse response of damped mass-

Fig 4: Power spectrum of helmet perturber

The head perturber is a crucial element of the testing apparatus. The torque perturbation created by the head perturber is one of the two controlled inputs for the experiment, the trajectory of the visual target being the other. The torque signal's amplitude and frequency distribution must be properly shaped to test the vestibular system across its entire bandwidth, from below 0.5 Hz to above 20 Hz. In the first prototype system, a small servomotor, installed on the top of the helmet, creates the torque perturbation by moving the external ring about the test subject's head. The servomotor accepts a position input from the computer, and accelerates the ring to the new position as quickly as it is able. Since the servomotor has no torque or velocity control, little can be done to shape the resulting reaction torque. As previously discussed, when the helmet perturber is directed with a pseudo random binary sequence, which

should approximate white random noise, the power spectrum of the resulting torque is too limited for our vestibular testing needs. The helmet perturber's torque power spectrum is not significantly altered by changing the frequency at which the PRBS is executed or by adjusting the allowable travel of the aluminum ring – the only parameters that can be controlled on the helmet perturber. The servomotor's torque cannot be adjusted to maintain constant acceleration as it moves to its desired position; the motor accelerates to a maximum speed and maintains this speed until the desired position is reached. We are therefore unable to create long duration (low frequency) torque perturbations because reaction torque is only felt while the ring is being accelerated. It is particularly important to generate low frequency inputs, for that is where many of the interesting vestibular pathologies are seen, and where interactions between the vestibular and optokinetic systems occur.

An alternative perturber is being developed that will allow adequate shaping of the reaction torque, and will provide a means of generating low frequency torques. Rather than relying on a central motor that drives an external ring, the new perturber uses Lorentz force actuators to accelerate a liquid metal circumferentially about the test subject's head. By driving large currents through the liquid metal perpendicular to a permanent magnetic field, a force is generated in the third axis that accelerates the liquid metal around a channel. The force applied to the liquid metal is controlled by controlling the current running through the liquid metal. A low frequency torque can be developed by continuously accelerating the fluid in the channel. The reaction torque felt by the test subject is dependent on the change in momentum of the fluid, not just the force applied to the fluid. The change in momentum of the liquid metal will be velocity dependent since the net force acting on the liquid will be the difference between the Lorentz force and the viscous shear forces. A prototype of this new actuator has been built and used for initial feasibility studies (Figs 5 and 6). Instead of using a liquid metal, as in the final design, the prototype drives a solid copper ring around a channel. A 130 mm radius, 3 mm wide, 70 mm high copper ring sits inside a housing machined from cast nylon. The inner radius

of the housing is large enough to fit around an average head, and allows the housing to be attached to head attachment devices such as a helmet or pair of headphones. Four pairs of neodymium-iron-boron permanent magnets provide a 0.4 Tesla magnetic field across sections of the copper ring. Current is delivered to the copper ring through four pairs of graphite brushes in sliding contact with the copper ring at the top and bottom. Threaded cylinders allow the contact force between the ring and the brushes to be adjusted. The pairs of brushes and magnets create four distinct Lorentz force actuator modules.

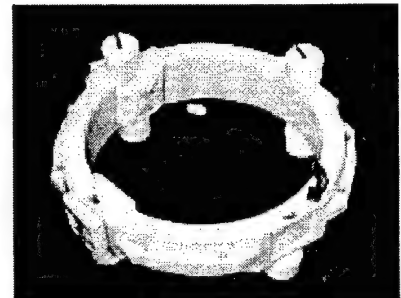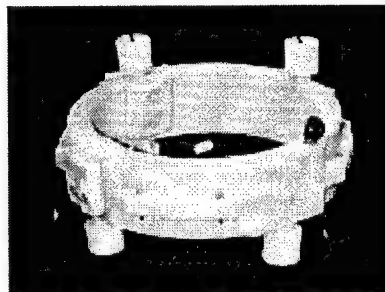Fig 5: CAD drawing of perturber housing

Fig 6: cast nylon housing for copper ring and liquid metal

We have found 1 N-m to be an appropriate level of torque for perturbing the head during the vestibular tests. To produce this magnitude of torque, the copper ring perturber requires over 100 amps of current at each of the four actuator modules. For the test prototype, 12-volt lead-acid batteries, wired through automobile ignition switches and 12 awg equipment wire were used to provide the necessary current. Numerous high current/low voltage power supplies are available on the market that can be used for the final apparatus. Testing consisted of quickly applying current to the copper ring and monitoring ring movement. We limited the time the current was passing through the ring to 1 second to reduce the risk of a fusing the ignition switches and creating a dangerous low resistance closed circuit with a battery. During testing, currents of up to 65 amps were recorded flowing through the wires and brushes of each actuator module. After a few 1-second on-off cycles the wires became very hot to the touch, but not hot enough to be in danger of fusing. Enough force was developed by the Lorentz actuators to just move the copper ring during the 1 second application. Friction between the ring and the

cast nylon housing is high and limits the acceleration of the ring. Smaller, lighter rings, were also tested, and moved significantly farther than the original copper ring. However their smaller width reduced the effectiveness of the brushes, and limited the current flow through the copper ring. Rather than building a bearing system to reduce the friction between the copper ring and the cast nylon housing, the solid copper ring is being replaced by a liquid metal. The copper ring was useful for a successful feasibility study, but more will be learned from a liquid metal system. A gallium-indium-tin alloy with a melting point of 11 c is being procured from a specialty metals company. This is one of the few non-toxic alloys that is in liquid form at room temperature. Using an alloy with a below-room-temperature melting point significantly reduces concerns about thermal expansion, filling the actuator, or heating before use. The gallium-indium-tin alloy's conductive and thermal expansion properties appear acceptable for our application, but little other data, including magnetic properties, are available from the manufacturer. Gallium-indium and gallium-tin alloys from the same company have acceptable magnetic properties, so it is hoped that the same will be true for the gallium-indium-tin alloy. The change in momentum of the fluid must be maximized to impart the greatest torque to the head, and so the fluid channel must be designed to limit the effects of the boundary layer in retarding the fluid flow. CFD analysis is being employed to optimize the channel's shape.

Although the general shape of the housing will not be changed, a few modifications will be necessary for the liquid metal. The graphite brushes will be replaced by copper strips that extend slightly into the liquid metal flow, and the square permanent magnets will be replaced by thinner, longer magnets.

The eloctro-oculograph (EOG), used to track eye position, has been completely redesigned (Fig. 7). It is still based on the signal conditioning board that was designed in-house, but the new EOG has been improved with added input/output channels, additional gain settings, low and high pass

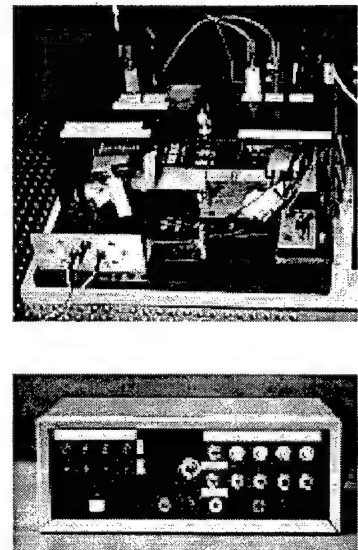Fig 7: Original and new electro-oculograph

filtering, and packaged in an electronics enclosure. The EOG can now

7

accept four channels of differential eye input, set gains of 200, 600, and

800 x, and has two channels with high pass filtering set at 0.07 Hz.

We now have the option of tracking individual eye motions in the vertical and horizontal planes, as well as the conventional horizontal binocular motions. A common problem with DC electro-oculography is signal drift. It is not uncommon to begin an experiment with the EOG signal centered at 0 vdc, and within a few minutes to develop a gradual 4 of 5 volts signal offset. When EOG gain is set to a moderate 500 x, it only takes a 0.01 volt difference between the reference electrode of the forehead and the electrodes attached to the outer canthus of the eye to create a 5 volt offset. At times the drift would drive the EOG signal beyond the +/- 10 vdc input range of the data acquisition board, making it impossible to collect accurate data. To help remedy signal drift, two of the four EOG channels are equipped with 0.07 Hz high pass filters. The filters effectively eliminate the signal drift, but the disadvantage is that the dc eye position is lost from the EOG signal. The dc position is important to some evaluations of gaze orientation, and can be used for clinical assessment of the oculomotor system. However, it is not essential to the development of our vestibular testing system, and can be solved later with a more costly commercial EOG or by further modifying our signal conditioning board with biasing electronics. For our system, the high pass filtering provides an acceptable, cost effective solution, and we are able to collect a clean eye signal with the present arrangement.

Experimentation with test subjects, with no known vestibular deficiencies, has been an important tool in the development of our equipment and software. The test subject is seated 1.45 meters from a flat paper screen on which the visual target is projected from a low power (100 mW), 650 nm wavelength laser. The target is adjusted so that it is projected at a height equal to the test subject's eyes. The target can be controlled either by a frequency generator for simple sinusoidal motion, or computer controlled to follow patterns that are less predictable by the test subject, such as a smooth sum of sine trajectories or with intermittant jumps from point to point. The smooth trajectories are designed to elicit slow phase nystagmus eye motions, while the faster, intermittant, target motions elicit fast phase or saccadic motions from the test subject's eyes. The helmet perturber is secured to the test subject's head with a chin strap, and a rotational velocity sensor is mounted to the bridge of the nose. The test subject is asked to

track the the target while the helmet perturber shakes their head. They are asked to track targets with natural head and eye motions, but to limit body translations, as the head sensor only records rotational motion. The helmet is tolerated well by most test subjects when moderate to small perturbations are delivered. This translates to executing the pseudo random binary sequence between 10 – 30 Hz, and allowing less than 10 degrees of travel of the external ring. Increasing the tarvel of the ring, and slowing the frequency at which the PRB sequence is executed, seems to cause slight nausea after in some test subjects. None of the test subjects have reported being unable to track the target while under the influence of the perturbations, though saccades are often visible in the EOG. The saccades indicate that smooth pursuit of the target was not always possible, and fast saccades were necessary to re-acquire the target. Individual trials lasted only 20 seconds, and test subjects were used for five to ten trials. Individual trials were tested at different helmet perturbation levels and visual target trajectories. Data were collected at 100 Hz on eye position, target position, rotational head velocity in the horizontal plane, and force applied to the test subject's head. Both velocity and position signals are needed for head, eye, and target motions, and since in each case only one of the two signals is collected, the other must be calculated. Calculating head position from rotational velocity is complicated only by a signal drift that is apparent when taking the integral of the signal, but it is easily removed. Determining clean velocity signals by approximating the derivative of the eye and target position signal is considerably more difficult. The velocity signal often displays spikes and other noise that occurs when differentiating a discrete time signal. Smoothing the signal, so that it may be analyzed requires both median and low pass filtering. Simple median filtering is effective at removing aberrant data points from the velocity signal, and low pass filtering between 15 and 20 Hz is effective at removing the higher frequency noise. During tracking, both slow and fast phase nystagmus are often seen in the eye motions, but only the slow phase nystagmus is used in the gain-phase analysis of the vestibular system. The fast phase nystagmus signal components, as well as the corresponding head, target, and force signals must be removed before using the signals to assess the vestibular system. A nystagmus classifier has been developed by Dr. Galiana of McGill University. Originally designed for Matlab, the classifier will be modified to be used with our Mathcad analysis software.

Drowsiness/Alertness Monitor

The alertness of a human operator is essential for the safe operation of automated machinery and vehicles and the accurate monitoring of alertness and the detection of drowsiness is critical to such functions as air traffic control, nuclear power plant operation and vehicle control. Industrial inspection tasks, especially those that are system rather than operator paced, involve perceptual processes that are particularly susceptible to the level of alertness in the human operator. It would appear that in this context monitoring alertness would be of considerable benefit, particularly in view of the results from a number of studies of detection performance in industrial inspection tasks which show high omission rates for flaws (average of 30%) in most of the products inspected which range from electronic equipment to foods and industrial products.

The Drowsiness/Alertness monitor that will be described here has been formally submitted as a patent (09/146,828) to the U.S. Patent and Trademark Office. The index that is being proposed to measure alertness is based on a number of variables including the frequency of saccades, gaze stability, saccade speed, blink duration and on the motion of the head as reflected in scanning the environment. A physiological indicator will be derived from these measurements whose value would define a state on the alertness-drowsiness continuum which could in turn trigger an alarm.

Studies of ocular reflexes indicate that under conditions in which the head is unconstrained and free to move, it is quality of gaze stabilization on targets in space, coupled with the speed with which new targets are acquired visually that determines the quality of ocular reflexes. Gaze stability encompasses both the stability of the eye within the head and of the head in space. When an individual becomes drowsy, gaze stability between saccades begins to drift and the number and peak speed of reorienting saccades decreases which results in the eye blink dynamics becoming sluggish. These changes in dynamics are reflected in changes in the vestibular-ocular reflex (VOR) which is measured quantitatively using the vestibular-ocular testing device. From measurements made of eye and head movements, the number and duration of saccades and blinks made by the operator can be determined as well as the gain of the VOR.

From these data, threshold values will be used to trigger alarms which may be auditory or visual to alert the operator or supervisor to a change in arousal state.
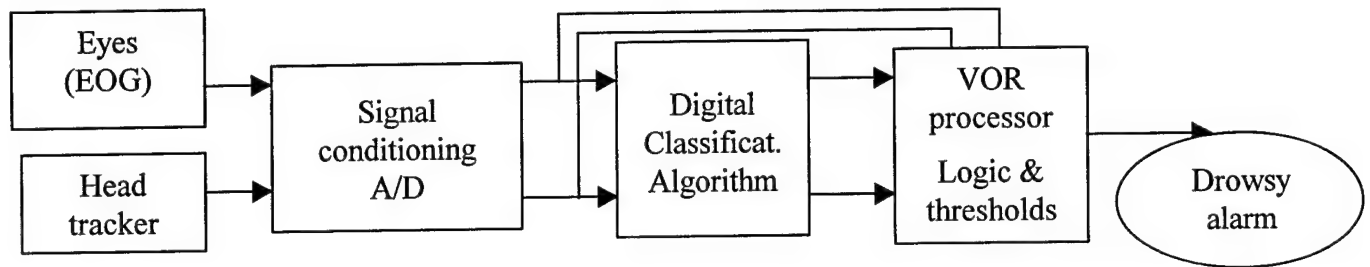
Figure 1: A schematic diagram of the main elements of the Drowsiness/Alertness Monitor

As shown in Figure 1, the motion of the head is tracked using either an inertial based sensor that has a bandwidth exceeding 20 Hz (e.g. the VOR testing apparatus) or a magnetically-based sensor (such as the Polhemus or Ascension Technologies devices). The position of the eyes is tracked using the electro-oculargram (EOG) which provides a signal corresponding to the direction of regard of one or both of the eyes. These two sets of data are digitized and filtered by the signal conditioning unit and then used by a VOR processor to calculate the ideal conjugate VOR gain taking into account the vergence, where the vergence set point is the sum of the two eye angles referenced to the nasal direction. The gain of the VOR is usually greater than or equal to 0.6 when viewing targets in the far field even in the dark. During drowsiness the gain of the VOR can decrease significantly. The analysis of ocular responses includes automatic classification of the slow and fast phases of eye movement based on classification software developed by Rey and Galiana (1991), and separately identifies blinks. The processor corrects for vergence and evaluates the stability of the gaze during the slow phases, and may also calculate the time spent by the operator viewing targets in different parts of the visual field.

Wearable Infrared Tympanic Thermometer

The core temperature of the human body exhibits one of the most stable of all 24-hour rhythms with a variation of about 0.5°C during the day. The peak temperature occurs around 19:00-20:00 hours and the trough occurs around 4:00-5:00 hours for people engaged in a normal lifestyle (i.e. sleeping 7-8 hours at night). The form of the circadian variation in body temperature is quite stable and markedly endogenous in origin, but is independent of the sleep/wake cycle from which is can be dissociated when external cues from the environment are no longer available (i.e. light/dark cycle, temperature variation).

The monitoring of body temperature over long time intervals could serve two purposes: first, the circadian variation in temperature could be correlated with human performance measures and used as a factor in determining the level of alertness of a human operator; second, body temperature is generally assessed to assist in the diagnosis of disease by detecting fever. Body temperature can be assessed from at least four sites: the axilia, the mouth, the rectum and the ear canal. The latter site has the advantage that it is easily accessible and that the measurement process does not interfere with the activities of the individual being monitored and requires little patient cooperation. An infrared tympanic thermometer is therefore being developed as part of the SIMSUIT project and is initially based on the design of the Thermoscan thermometer. The infrared sensors being considered for detecting infrared radiation from the tympanic membrane in the auditory canal are pyroelectric elements such as those used commercially in the Thermoscan and E-Z Therm thermometers and thermopiles which were first used in the FirstTemp (introduced in 1986 by Intelligent Medical Systems) and Genius (introduced in 1991) thermometers. The infrared sensor chosen for the wearable infrared thermometer will be selected on the basis on a number of considerations, including cost, reliability, accuracy, speed of response, low noise and susceptibility to fluctuations in ambient temperature. In general, infrared tympanic thermometers are susceptible to errors if the ambient temperature is outside a specified range which is usually between 15-40° C. Although thermopiles used as thermal sensors demonstrate good accuracy and are reliable with low noise, they are seriously limited by the low level of output signal, non-linearities and relatively high cost. In contrast, pyroelectric sensors such as those used in the Thermoscan thermometers are heat flow detectors and so only measure a change in temperature rather than an absolute temperature. For this reason, pyroelectric sensors must be used with both a mechanical shutter

and an internal reference sensor such as a thermistor to measure ambient temperature. Core body temperature as measured by the temperature of the blood perfusing the thermoregulatory receptors in the hypothalamus is then calculated from the difference between the temperature detected by the pyroelectric sensor when the shutter is opened and the ambient temperature recorded by the thermistor.

The infrared tympanic thermometer that is being designed incorporates a probe in the ear canal that directs the infrared radiation from the thermal target to the infrared sensor and a sensor unit, which comprises an infrared detector, and a signal-processing unit that converts the thermal radiation into an electrical signal. The signal processing unit and power supply will be mounted in the headset that the person wears and the probe will protrude from one of the earphones. Pyroelectric sensors will not be used because of the necessity of using an actuated shutter which requires moving components.

References:

Rey, C. & Galiana, H.L. (1991). Parametric classification of segments in ocular nystagmus. IEEE Transactions on Biomedical Engineering, 38, 142-148.

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Patient Monitoring and Diagnosis

CHAPTER 5

An Intelligent Cardiopulmonary System for the Home Health Market
T. Sheridan

AN INTELLIGENT CARDIOPULMONARY SYSTEM FOR THE HOME HEALTH MARKET.

Thomas B. Sheridan

HYPOTHESIS

An intelligent cardiovascular diagnostic system (ICDS) will improve patient care and decrease medical costs by the earlier detection of patient decompensation.

1. INTRODUCTION

The goal is this project is the development of a system designed to acquire, process and analyze blood pressure, heart rate, oxygen saturation, and thoracic signals to make a decision on the relative health of the patients cardiovascular system. Unfortunately most patients do not have the tools necessary to become an active and informed participant in their own health care. Many of those who end up with serious health problems enter the health care system too late, and thus require more extensive and costly care. The patients selected for monitoring by the intelligent system are those patients who are at a higher risk for decompensation as compared to the general population. These "high risk" patients frequently enter the health care system too late and thus require more extensive and costly care in addition to the emotional and physical strain to themselves and their families. The goals of this program are to decrease the initial acuity, length of hospital stay and readmission rates for patients with congestive heart failure. This will result in substantial savings in health care costs with a decreased burden on the acute health care system.

Why focus on the cardiovascular and pulmonary system? It is estimated that 65 of 239 million Americans have cardiovascular disease. One million die annually, and this is one of every two deaths in the United States. The mortality from cardiovascular disease exceeds that of all other diseases combined. Congestive heart failure (CHF) is estimated to affect

three million people in the United States. It is the final pathway of a variety of primary cardiovascular disease entities, such as coronary artery disease, hypertension, valvular heart disease, genetic disorders, diabetes and the sequelae of infection or toxin exposure, among others. Hospitalizations and mortality from CHF have increased steadily since 1968, despite the overall improvement in mortality from cardiovascular disease. Heart failure is now the underlying cause of death in over 39,000 persons annually. In 1992, it was the first listed diagnosis in 822,000 persons and is the most common hospital discharge diagnosis in persons over 65 years of age. The incidence of death from CHF is 1.5 times as high in black Americans as in whites. The estimated direct economic cost of CHF in the United States be reported to be $10.2 billion annually. The problem will only get worse, as the elderly segment of the population is increasing at a rate 5.6 times that of the other age groups. There are currently 25 million Americans greater than 65 years of age and 2.7 million Americans greater than 85. Over the next 50 years the >65 age group will see a 140% increase versus 25% in the other age groups. At present, the only cure for end-stage CHF is cardiac transplantation.

Studies have shown that intervention can improve care and decrease costs by decreasing hospital admissions, which account for a large portion of their health costs. Investigators in Los Angeles found that interventions (invasive tests, medication adjustment, patient education and follow-up at a heart failure center) decreased the number of hospital admissions from 429 in the six months before referral to 63 in the six months after referral.[1]

2. PROPOSED SYSTEM

As stated earlier, the goal is this project is the development of a system designed to acquire, process and analyzed blood pressure, heart rate, oxygen saturation, and thoracic signals to make a decision on the relative health of the patients cardiovascular system. The patients selected for monitoring by the intelligent system are those patients who are at a higher risk for decompensation as compared to the general population. These "high risk" patients frequently enter the health care system too late and thus require more extensive

and costly care in addition to the emotional and physical strain to themselves and their families.

We propose to produce two types of intelligent systems. The emphasis will initially be placed on a portable system that will be carried by the home health care professional on a visit to the patients' home. A second system will be a permanent home based system for use by the primary caregiver and / or fragile (CHF) patient. In either scenario, the operator of the Intelligent Cardiopulmonary Decision System (ICDS) will be directed by the ICDS on where to place various sensors and what measurements to take. The ICDS will then process the data and make a recommendation to the patient concerning further care. There will be human factors issues on the user interface, as well as some type of patient feedback so that they can actively participate in their care.
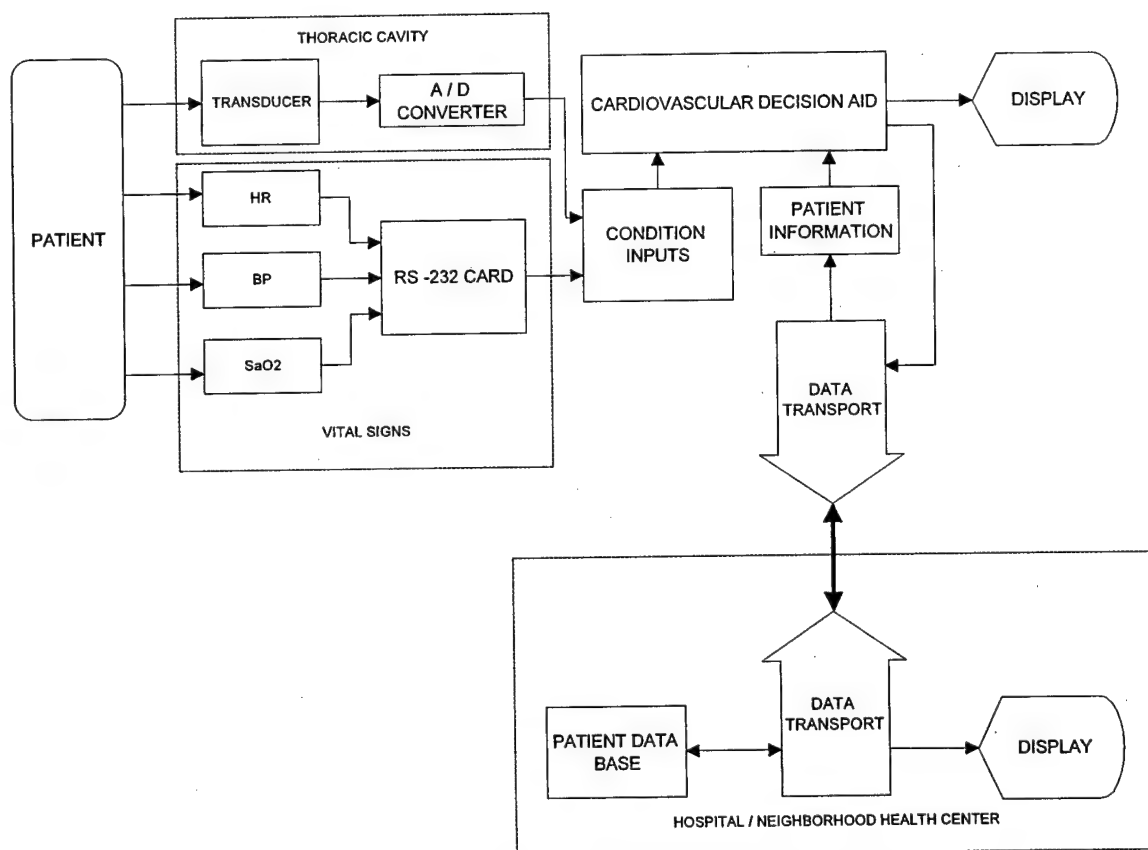

Figure 1. Structure of the Intelligent Cardiopulmonary Decision System (ICDS)

The structure of the ICDS is shown in figure 1. The patient's vital signs and oxygen saturation will be acquired first and will be evaluated by the ICDS. The ICDS will then direct the acquisition of other information as needed by the system. This additional information will be conditioned, digitized, and processed before being put in a form that could be inputted into the ICDS. The ICDS will then make an initial assessment of the patients current state, compare it with a predetermined "optimal state" and make a decision on where to proceed from that point. Options include requesting additional information from the patient, patient education, instructions to hold or to take an additional dose of a medication, decision to re-evaluate after a waiting period, contact the primary care physicians office, connect to the central system, or call an ambulance for transportation to a medical facility.

3. HEART RATE ANALYSIS MODEL

A system was developed to obtain the ECG waveform, as well as digital values for the oxygen saturation, systolic and diastolic blood pressure readings, and time averaged pulse information (Figure 2).
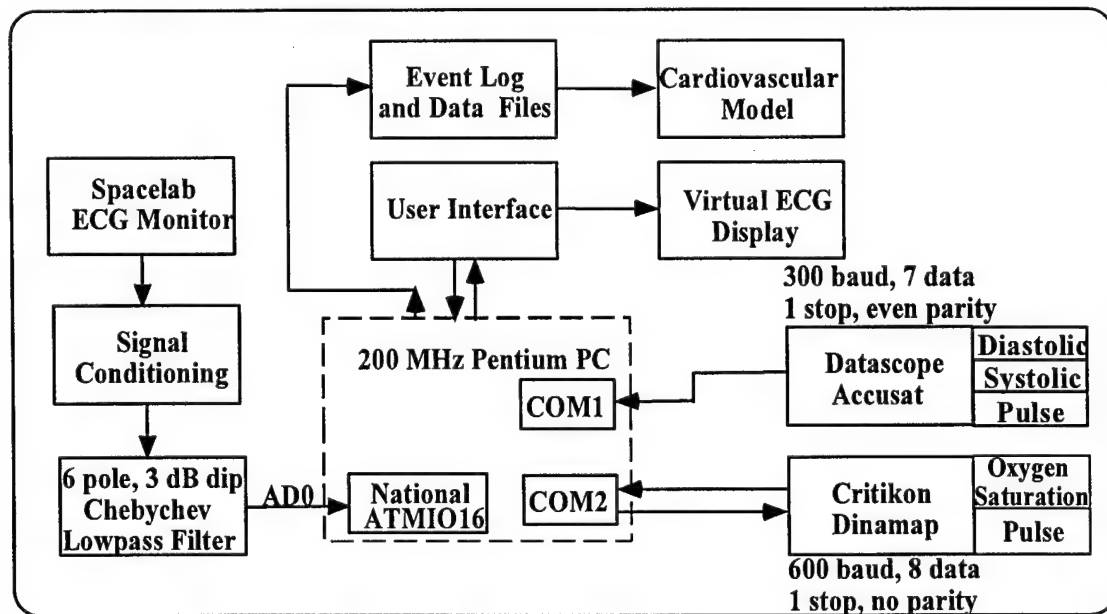

Figure 2. Heart Rate Analysis System

The distance in time between two consecutive R-waves of an ECG waveform defines the instantaneous heart rate. With the instantaneous heart rate known over a period of time, it is possible to find the variability of the heart. This is accomplished through the use of the power spectral density function, which gives insight into the efficiency of the patient's cardiovascular system. The developed system was tested on four healthy adults, with the resulting data for one patient presented here.

3.1 R-wave Peak Detection and Heart Rate Variability

Since the exact location of an R-wave peak is needed for heart rate variability studies, it is highly important to condition the analog ECG signal with not only the normal amplification, but also with an anti-aliasing low-pass filter. In order to produce a sharp falloff from the pass-band to the stop-band frequency range, a sixth order, 3-dB-dip Chebyshev low-pass filter was chosen. A cutoff frequency of 170 Hz was selected since it was above the maximum frequency component of the R-wave of the typical ECG waveform. Figure 3 shows a ten-second window of ECG data taken from one of the test subjects.
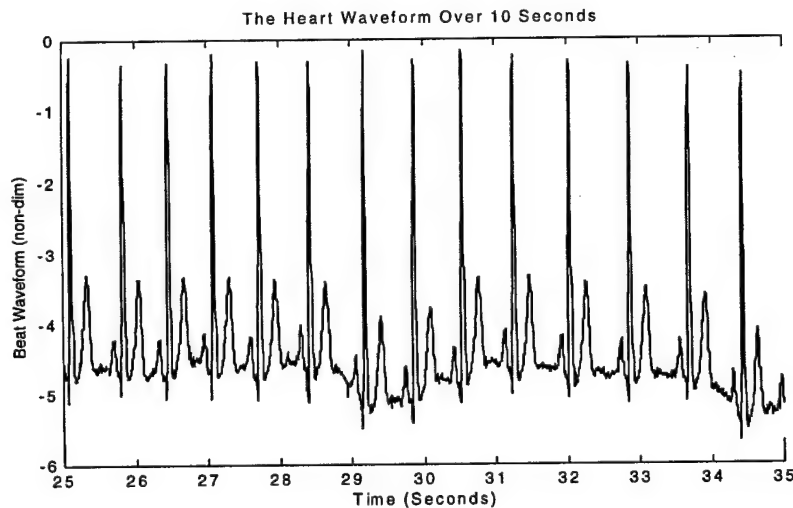

Figure 3. A Ten Second Window of the Sampled ECG Waveform

From this ECG waveform, R-wave peak detection was applied so that the interval of time between peaks could be used to produce an instantaneous heart rate waveform. The first

step in this procedure was to digitally filter, forward and backward, the ECG data. This forward and backward data was then summed together to eliminate the phase shift associated with digital low-pass filtering. Once the high frequency noise was removed from the ECG signal, a derivative was taken of this sequence. A "window" was then passed through this new sequence, g[n], with the magnitude of the derivatives summed over the window. This new sequence can be written as follows:

$$y[n] = \sum_{n-\Delta n_s}^{n+\Delta n_s} \left| \frac{g[n+1]-g[n]}{\Delta n} \right| \tag{3.1}$$

$$\text{where } \Delta n_s = \frac{n_1+n_2}{2}$$

Since the purpose of this window was to extract the R-wave peak, the time-width of this window was set to be of the order of the inverse of the R-wave frequency maximum:

$$\frac{1}{t_s(n_2 - n_1)} \sim f_{\text{R-wave}} \tag{3.2}$$

where t_s is the sample time of the analog input. The sample frequency was 500 Hz., and with 170 Hz for the maximum R-wave frequency component, a window three elements wide in discrete time was needed. Figure 4 shows the results of this windowing enhancement of the ECG R-wave peak.
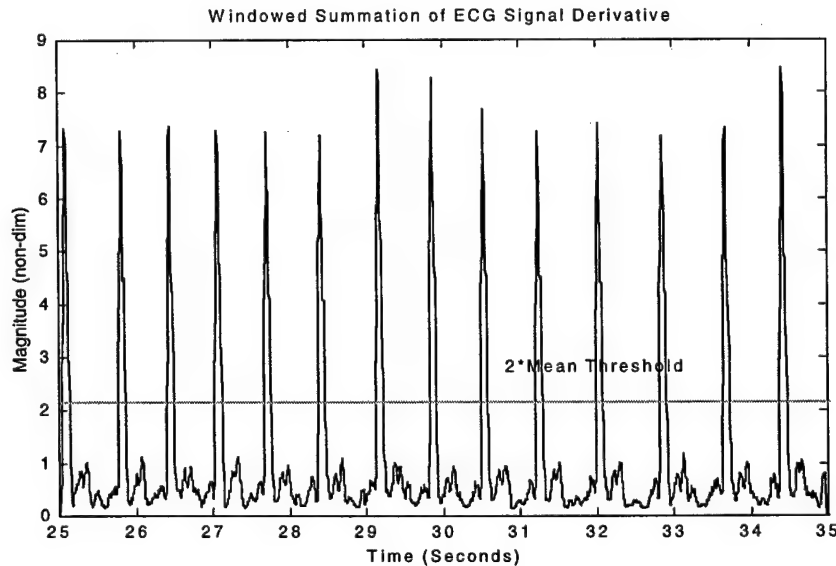

Figure 4. Enhancement of the ECG Waveform Peaks

6

Once the windowing enhancement was accomplished, the mean of the resultant sequence was calculated, and a two-mean threshold was set on the sequence for identification of the R-wave location. Figure 5 shows the after threshold sequence.
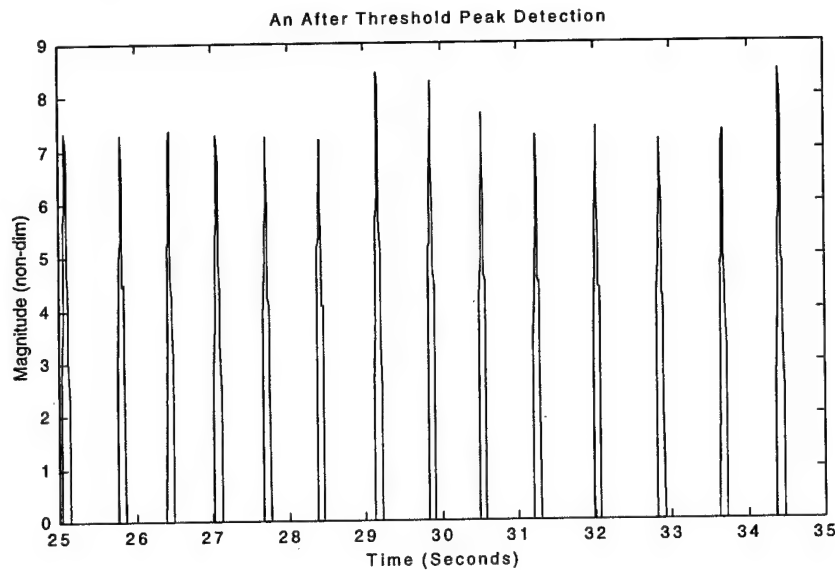

Figure 5. The After Threshold R-Wave Locations

It is important to note that since the R-wave changes in frequency makeup and magnitude between patients, in addition to a changing DC value of the ECG mean that may, at times, be above a set threshold, it is not possible to simply set a threshold on the ECG directly to determine the R-wave location. Figure 6 shows the location of the peaks found with respect to the original ECG data.
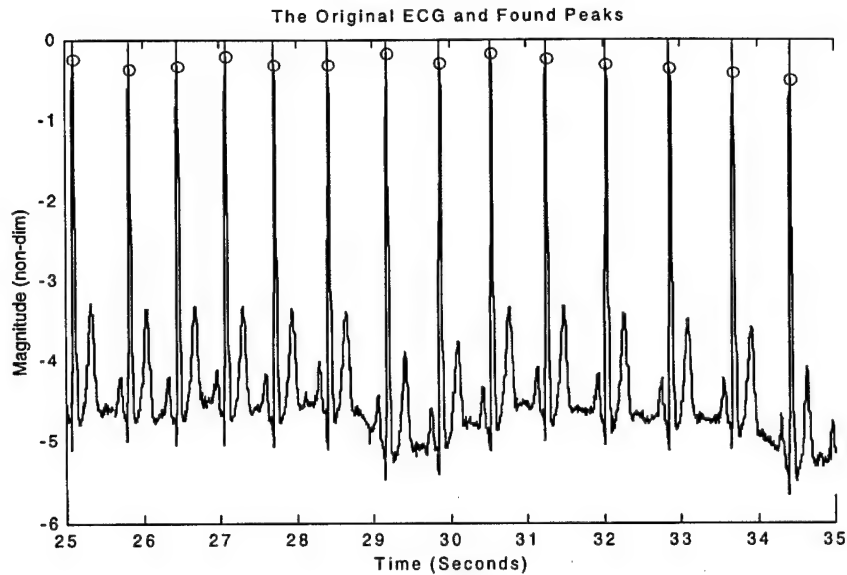
Figure 6. Peaks Detected Within the Original ECG Waveform

With the peaks of the R-wave located, the time between the peaks was calculated, and the inverse of this was the instantaneous heart rate. Figure 7 shows the instantaneous heart · rate over 10 seconds for one test subject. Since the human heart has beat variation, and beats at unequal increments of time, the instantaneous heart rate waveform is not a straight line. This is not the case with heart transplant patients, in which the heartbeat is well regulated, giving no beat variation, and producing a straight-line instantaneous heart rate waveform. This waveform, which has unequal time spacing between beats, is then sampled at a frequency that is roughly four times the maximum frequency of the instantaneous heart rate waveform. This gives a discrete representation of quasi-continuous data that can be used in variational analysis. After it is re-sampled, it is digitally low-pass filtered again, and then re-sampled again, which basically converts it from the true instan It used in produ g the R-wave id then a low-p method that v

8
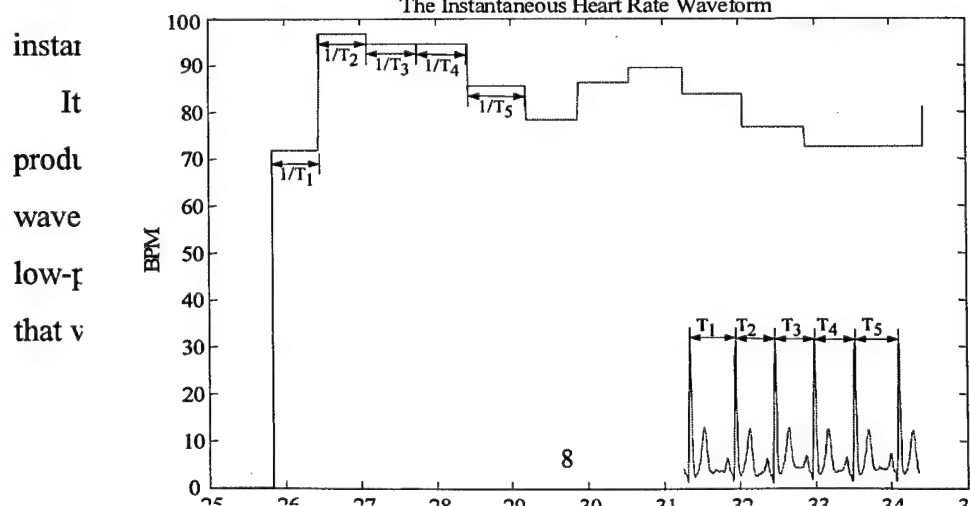
Figure 7. The Instantaneous Heart Rate Waveform

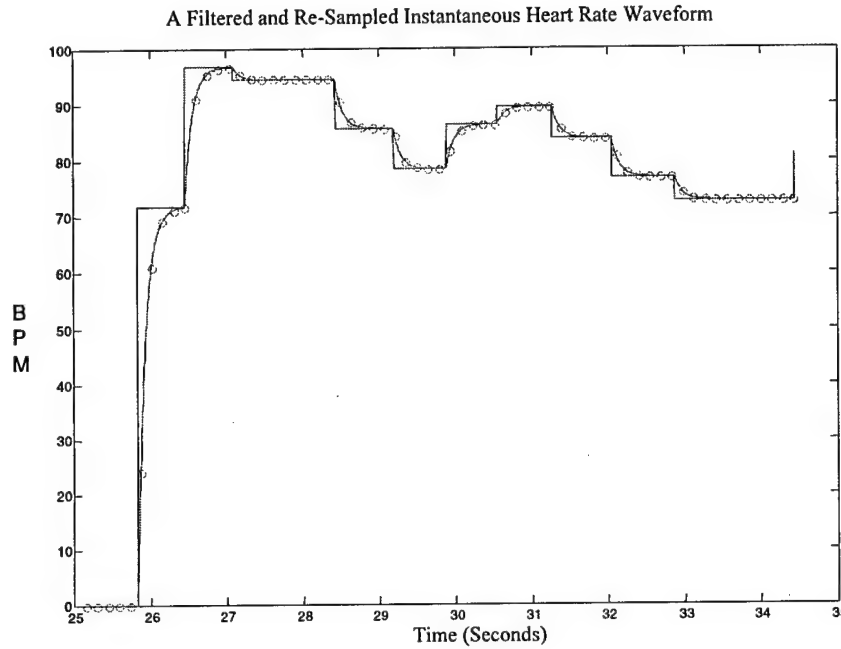A Filtered and Re-Sampled Instantaneous Heart Rate Waveform

Figure 8. A Re-sampled Instantaneous Heart rate Waveform

Although the above is only a ten second window to demonstrate a method, the four test subjects were put through several ten minute sessions in which ECG, oxygen saturation, systolic and diastolic blood pressure, and pulse information were monitored.

In order to measure the heart rate variability, a power spectral density for the instantaneous waveform had to be found. The first step to achieve this was to find the autocorrelation, $R_x[n]$, of an N-point sequence, $y[n]$ (the re-sampled heart rate waveform):

$$R_X[n] = \sum_{k=-N}^{N} y[k]y[n+k] \qquad (3.3)$$

The power spectral density (psd) function is defined as follows:

$$S_x(\omega) = \int_{-\infty}^{\infty} R_X(\tau)e^{-i\omega\tau}d\tau = \mathbf{F}\left[R_X(t)\right] \qquad (3.4)$$

For a continuous function g(t), the Fourier transform is:

9

$$G(\omega) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} g(t)e^{-i\omega t} dt \qquad (3.5)$$

From this it can be seen that the psd is nothing more than the Fourier transform of the autocorrelation function multiplied by a constant:

$$S_X(\omega) = \frac{1}{2\pi} \mathbf{F}\left[R_X(t)\right] \qquad (3.6)$$

Figure 9. The Instantaneous Heart Rate Waveform over a Ten Minute Period

With the instantaneous waveform known over the full ten-minute period, the power spectral density function was then found. Figure 9 shows this waveform.
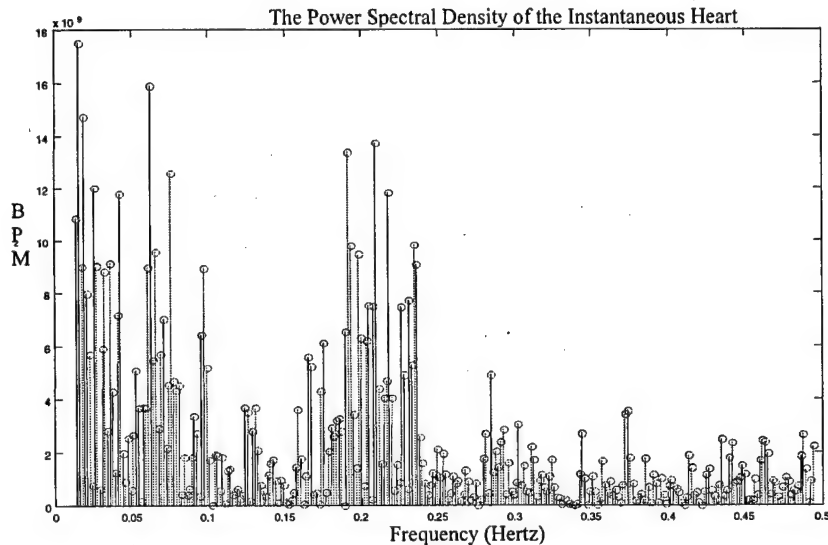

Figure 9. The Power Spectral Density Function of the IHR

There are a few important details to note at this point. First, there are roughly three peaks: one at roughly 0.03 Hz, 0.08 Hz, and 0.2 Hz. These are characteristic of healthy patients in variability studies. Second, the spectrum is not plotted below about 0.01 Hz, since the component below this frequency represents the mean drift of the ECG waveform, and has a rather large value. Third, the spectrum is not plotted above 0.5 Hz, since the values above this frequency are roughly zero.

2.2 More Information for the Cardiovascular Model

In addition to heart rate variabillity and the ECG information, diastolic and systolic blood pressure information as well as a time averaged pulse was brought in serially from a Critikon Dinamap that was set to automatically sample once a minute. Oxygen saturation and another time averaged pulse was also brought in serially, this time from a Datascope Accusat roughly once every fifteen seconds. This sample rate was limited by the embedded instuction set in each of the two pieces of equipment. Figure 10 shows this information over the ten minute sample taken for the patient.
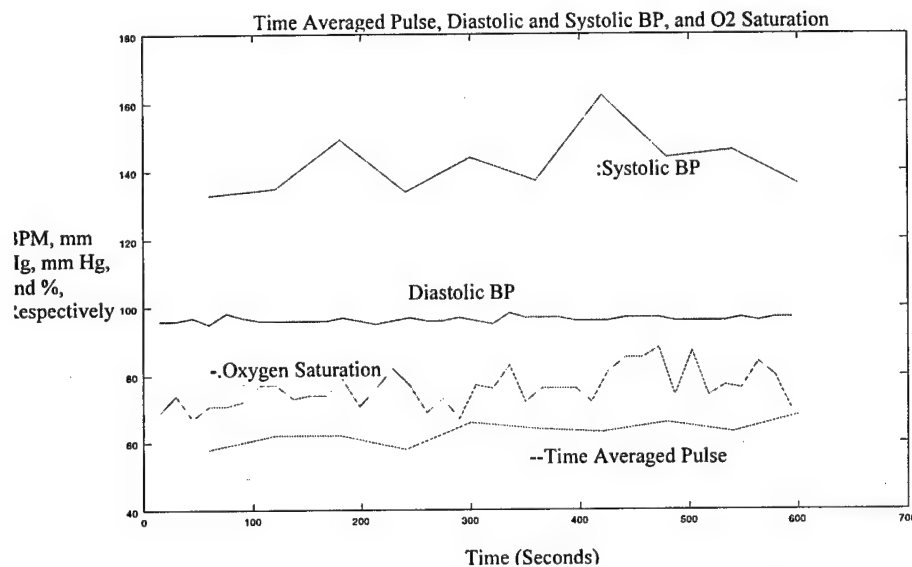
Time Averaged Pulse, Diastolic and Systolic BP, and O2 Saturation

Figure 10. More Patient Data for the Model

Heart rate is important in patients who currently have marginal coronary flow and are sensitive to the physiological consequences of tachycardia. Tachycardia decreases coronary diastolic filling time, which decreases the supply of oxygen to myocardial tissue, especially endocardial. In addition, tachycardia increases oxygen demand, which further contributes to negative myocardial oxygen balance. This initially results in regional wall motion abnormalities, which causes a rise in ventricular end diastolic and end systolic pressures, which further decrease diastolic blood flow, starting the cycle to heart failure. Bradycardia can also have deleterious effects on certain pathologic states. Patients with mitral or aortic regurgitation can go into congestive heart failure, depending on the magnitude of the regurgitant fraction and the degree of bradycardia. Changes in the other inputs would affect the magnitude of the changes in heart rate that would start the cycle

11

toward CHF. Both an increase and a decrease in blood pressure can have an effect on cardiovascular dynamics that would have a deleterious effect on cardiac patients. Certain types of congestive heart failure are sensitive to changes in afterload, and the presence of blood pressure changes in these patients could start the process toward congestive heart failure. Oxygen saturation, in addition to the heart rate variability, will be used by the model to quantitate the direction / presence of cardiac decompensation.

---

[1] Fonarow GC et al. *Impact of a comprehensive heart failure management program on hospital readmission and functional status of patients with advanced heart failure.* J Am Coll Cardiol 1997 Sept; 30:725-32.

Tissue modification with feedback: the "Smart Scalpel"

E. L. Sebern, C. J. H. Brenan, and I. W. Hunter

Department of Mechanical Engineering, Massachusetts Institute of Technology

Cambridge, MA 02139

ABSTRACT

While feedback control is widespread throughout many engineering fields, there are almost no examples of surgical instruments that utilize a real-time detection and intervention strategy. This concept of closed loop feedback can be applied to the development of autonomous or semi-autonomous minimally invasive robotic surgical systems for efficient excision or modification of unwanted tissue. Spatially localized regions of the tissue are first probed to distinguish pathological from healthy tissue based on differences in histochemical and morphological properties. Energy is directed to only the diseased tissue, minimizing collateral damage by leaving the adjacent healthy tissue intact. Continuous monitoring determines treatment effectiveness and, if needed, enables real-time treatment modifications to produce optimal therapeutic outcomes. The present embodiment of this general concept is a microsurgical instrument we call the Smart Scalpel, designed to cause hair growth delay or permanent hair removal.

1. BACKGROUND AND MOTIVATION

1.1 Feedback control

Feedback control (Figure 1) is widespread throughout many engineering fields, such as manufacturing, robotics, and in other human-machine interfaces. Feedback control uses measurement of the system output to modify the input in real-time. This on-line measurement strategy is necessary due to the difficulty of creating a comprehensive model to accurately predict the system output based on the input parameters. Feedback control provides a means to quickly respond to changes in the physical system or perturbations in the environment.
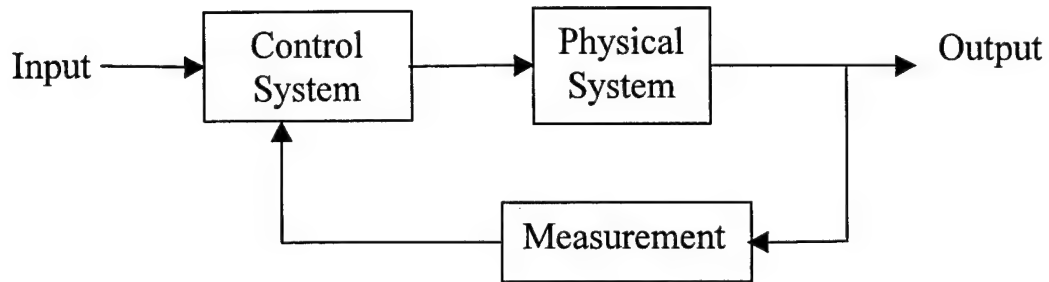
Figure 1: Illustration of classical feedback control loop in which the control system uses real-time measurement of the output to control the input to the physical system.

1.2 Application to medicine- the Smart Scalpel

An interesting application of feedback control is in the field of microsurgery. Many microsurgical procedures require a high degree of physical dexterity, accuracy, and control, which degrades rapidly with physician fatigue. This problem could be partially alleviated through inclusion of low-level decision-making embedded in a microsurgical instrument to aid in tissue location and removal. Our embodiment of this concept is an instrument we call the Smart Scalpel (Figure 2).
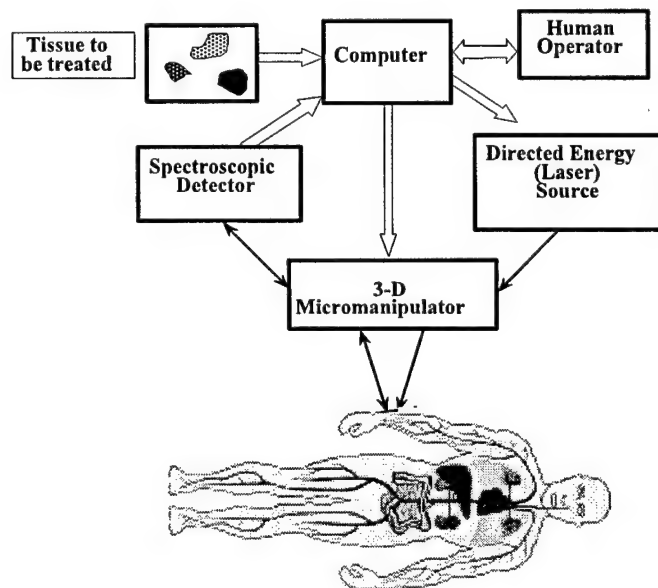

Figure 2: Schematic illustration of the "Smart Scalpel."

Implementation of the Smart Scalpel design is quite general in both measurement and intervention techniques. The human operator and computer model supply the computer with information regarding the properties of both normal and diseased/damaged tissue. The computer compares this information with real-time feedback of the histochemical and morphological properties of the tissue. Possible measurement techniques include: optical reflectance spectroscopy, magnetic resonance imaging, fluorescence optical spectroscopy, Raman spectroscopy, optical polarization detection, fluorescence polarization detection, mechanical impedance measurements, and electrical impedance measurements. The computer uses this feedback signal to identify the targets for a directed energy source, which affects only the diseased or damaged tissue and leaves the healthy tissue intact. The Smart Scalpel can employ a wide range of directed energy sources including: photon beam, electron or proton beam, localized electrical field, directed acoustic energy, and inertial cutting (low frequency mechanical energy). A micromanipulator serves as the interface between the patient and the imaging/therapeutic systems.

The many desirable attributes of the Smart Scalpel have the potential not only to improve the surgical performance in current microsurgical procedures but may yield the possibility of new surgical procedures not yet feasible with existing technology. The accuracy and reliability of present-day procedures may be enhanced and collateral damage minimized through an effective combination of analysis to discriminate between tissue types (e.g. diseased versus healthy) and removal/modification of the targeted tissue with a directed energy source. The Smart Scalpel diagnostics provide quantitative, rapid, on-line assessment of the procedure efficacy. This system of real-time feedback has great potential to increase patient comfort, shorten patient recovery times, and decrease the overall cost per procedure. Additionally, the Smart Scalpel is amenable to integration into a tele-operation system for remote surgery.

2. SMART SCALPEL APPLICATION TO HAIR REMOVAL

2.1 Current Clinical Practice

One application for the Smart Scalpel is in the area of permanent hair removal. Current methods for temporary hair removal include: shaving, cold or hot wax epilation, and chemical depilatories that often cause contact dermatitis.[1] Electrolysis is a permanent hair removal technique, but this method is tedious and only partially effective. Regrowth rates of 15% to 50 % have been associated with electrolysis.[2]

Because these hair removal techniques do not produce optimal therapeutic outcomes, laser hair removal has been explored. Two methods of laser hair removal have been tested with varying success. The first makes use selective photothermolysis to destroy or damage hair follicles using a normal-mode ruby laser (694 nm, 100-600 kJ/m^2, 270 µsec pulse width, 5-10 mm diameter spot). There are two dominant chromophores in skin that absorb this laser energy, melanin and oxygenated hemoglobin. Melanin is a chromophore present in the hair shaft or follicles, or both, which is absent in the dermis surrounding these follicles.[3] In the band from 650 nm to 700 nm, melanin absorption strongly dominates oxyhemoglobin absorption.[4] These longer wavelengths also penetrate deeply into the skin (550 µm to 750 µm) to selectively heat hair follicles in the underlying dermis.[5] Six month and two year follow-up studies of this normal-mode ruby laser treatment revealed that of the thirteen subjects tested, all laser exposures caused a hair growth delay, while four of the thirteen caused permanent hair removal of more that 50% of the treated region. The mechanism of laser hair removal is not well-understood, but histological studies from the normal-mode ruby laser study showed permanent hair removal correlated with miniaturization of the terminal hair follicles, rather than complete destruction of these structures.[6,7]

A second laser-based method of hair removal requires application of a carbon particle suspension, which fills the hair follicles, to selectively absorb energy from a Q-switched Nd:YAG laser. Mean percentage of hair regrowth at 1 month was 39.9%. At three months, the percentage of hair regrowth approximately doubled. The conclusion was that a single hair-removal treatment with the Q-switched Nd:YAG laser is safe and effective in delaying hair growth for up to 3 months.[8]

These laser treatments have drawbacks. In the normal-mode ruby laser treatment, great amounts of energy heat the skin regions without hair, making the procedure painful if the skin is not cooled. Currently a transparent material such as sapphire or glass is necessary to remove this heat from the epidermis.[3] Hair removal with both methods is not completely permanent. In the majority of cases, laser treatment only delays hair regrowth.

2.2 Smart Scalpel Approach
The more efficient Smart Scalpel approach is to first identify the hair follicles and target the laser to heat only these structures. This strategy leaves the tissue surrounding the follicles intact, so collateral damage is minimized. It is possible that by focusing laser energy on individual follicles, the Smart

Scalpel will be more effective in permanently removing hair, rather than delaying hair regrowth. Implementation of the Smart Scalpel strategy requires two elements: (1) a method to locate the hair follicles within the skin and (2) a means to direct the heating laser beam to the appropriate targets.

2.2.1 Spectroscopic identification of hair

Our approach to identify hair follicles makes use of selective absorption of light by the melanin present in the hair shaft and follicles. As mentioned earlier, melanin and hemoglobin are the two dominant chromophores of skin. A visible light reflectance spectrum of whole blood reveals high reflectivity of hemoglobin near 650 nm to 700 nm.[5] Melanin absorbs at this wavelength. Since melanin is present in the hair shaft and follicles, we can illuminate the follicle with 650 to 700 nm light and distinguish the absorbing hair follicles from the highly reflective blood vessels. Hemoglobin has a higher relative absorbance from 520 nm to 580 nm. Therefore, by taking the ratio of skin images illuminated with melanin-absorbing and hemoglobin-absorbing wavelengths, the hair signal can be clearly distinguished from the surrounding tissue. Melanin content in a dark human hair is 2.32% by weight, while human skin contains 0.023% and 0.008% by weight for dark and light-skinned patients, respectively.[9] Therefore, the strong melanin absorbance signal from hair can be distinguished from the melanin signal of other skin structures. Some image processing may be required to differentiate between the follicle and the rest of the hair.

2.2.2 Laser source

Once the follicles are identified, an appropriate laser source must be used to destroy these targets. We have identified several specifications for the laser source related to fluence, beam diameter interacting with the skin, laser size, and time-scale of laser-tissue interaction. Fluence levels depend on the wavelength of the laser used. In current treatments at 694 nm, the laser fluence is in the range of 100 to 600 kJ/m^2.[6,7] The wavelength of the laser light determines the coupling efficiency, which is the amount of energy applied to the tissue that is converted to thermal and/or mechanical energy. Therefore, if we use a laser wavelength, such as 1064 nm that melanin does not absorb as readily, we must increase the fluence of our laser.

We plan to illuminate the hair follicles with a 20 μm diameter laser beam, ~1/10 the diameter of a hair follicle. The common clinical practice in laser hair removal is to use a 5- 10 mm diameter beam.

Therefore, we can achieve the same fluence with laser energies $\sim 10^5$ times smaller than currently needed to destroy the follicles. A laser with a single spatial mode allows us to focus this beam to the small diameters required. Finally, our pulse width will determine whether the laser-tissue interaction is thermally-mediated or an adiabatic, mechanical process; this will be further developed in the following section. The size specification requires that the laser be small and lightweight so the physician and patient can comfortably interface with the SmartScalpel. If the laser cannot be made compact, we must be able to transmit the laser energy through a fiber optic so the laser is remote from the SmartScalpel.

2.2.2.1 Thermal Relaxation Time

In general, laser-tissue interactions can be grouped into two broad thermodynamic categories based on the time scale of the interaction. The dominant thermodynamic regime is characterized by the ratio of the laser illumination time, τ_{ill}, to the tissue thermal relaxation time, τ_r. For $\tau_{ill} \geq \tau_r$, the illuminated region is in thermal equilibrium with the surrounding tissue, and the tissue removal/transformation is primarily thermally mediated. When $\tau_{ill} < \tau_r$, laser energy is absorbed faster than it can be transported away from the illuminated region, and adiabatic processes determine the partitioning (and ultimate dissipation) of energy in the affected tissue volume. The thermal relaxation time for a cylindrical object, like a hair follicle, is given by:

$$\tau_r = \frac{d^2}{16\alpha} \qquad , \qquad (1)$$

$$\alpha = \frac{\beta}{\rho \cdot c} \qquad , \qquad (2)$$

where d is follicle diameter, and α is the thermal diffusivity, which is a combination of β, ρ, and c, the thermal conductivity, density, and specific heat, respectively.[10]

For tissue composed of 70% water, the material constants have the following values:

$$\beta = 4.21 \times 10^{-3} \frac{cm^2}{s}$$

$$\rho = 1.09 \frac{g}{cm^3}$$

$$c = 3.35 \frac{J}{g \cdot K}$$

Therefore, the thermal diffusivity, α, is 1.15×10^{-3} cm^2/s.[10] Assuming a 200 μm diameter blood vessel, which is typical for ectactic vessels in a port wine stain, the thermal relaxation time is approximately 20 msec.

Present laser-hair removal utilizes each of these two different thermodynamic regimes. The normal mode ruby laser, with longer pulse widths of 270 μsec relies on thermally-mediated processes to destroy the blood vessels. Alternatively, the Q-switched Nd:YAG laser, with pulse widths on the order of nanoseconds or picoseconds heats the carbon particles in a shorter timescale, destroying and/or damaging the hair follicle via an adiabatic, mechanical mechanism.

2.2.2.2 Light Absorption and Scattering

As laser light propagates through tissue, its intensity is attenuated by absorption and scattering. Beer's Law describes the amount of light attenuated by a tissue:

$$\frac{P_{out}}{P_{in}} = e^{-\mu_t x} \qquad , \qquad (3)$$

where P_{in} is the power delivered to the tissue, and P_{out} is the power that passes through the tissue, in other words, the power that is not absorbed or scattered. The parameter μ_t is the optical extinction coefficient with the dimensions of [1/L], and x is the length over which the light interacts with the tissue. The mean free path is the value of x equal to $1/\mu_t$. Over the distance of one mean free path, the ratio of P_{out} to P_{in} is e^{-1}, so that P_{out} is approximately 35% of P_{in}. The optical extinction coefficient, μ_t, is given by:

$$\mu_t = \mu_a + \mu_s \quad , \qquad\qquad (4)$$

where μ_a and μ_s are the absorption coefficient and scattering coefficient, respectively.

The scattering coefficient determines how much of the light originating at the surface of the tissue actually reaches the structure of interest. When the characteristic dimension of scattering particles is much less than the wavelength of light passing through the tissue, Rayleigh scattering describes a relationship between scattering and wavelength. In this case, the power lost to scattering, P_s, is proportional to $1/\lambda^4$.

The absorption coefficient governs the amount of non-scattered energy that is absorbed by the structure,[11] which is a hair follicle in our application. The absorption coefficient varies with wavelength, and every material has its characteristic absorption spectrum. Melanin has much stronger relative absorption from 650 to 700 nm than oxyhemoglobin. Current normal-mode ruby laser therapy makes use of this selective heating for permanent hair removal. The absorption of melanin is much higher at shorter wavelengths, especially in the ultraviolet band, which would make shorter wavelength lasers more effective in heating the hair follicles. However, the amount of scattering at this wavelength is also substantially higher. The relative scattering for the 694 nm normal-mode ruby laser versus a 248-nm krypton-fluoride laser results in penetration depths of 750 μm and 2 μm, respectively.[5] Since hair follicles are approximately 500 μm below the skin surface, these longer wavelengths are required.

2.2.3 Microchip Laser

Given that the Smart Scalpel must comfortably interface with the physician and patient, the small microchip Nd:YAG laser, developed at MIT's Lincoln Laboratory may be a useful source for permanent hair removal. While the normal-mode ruby laser used in currently seems to yield the best clinical results, the strategy of focusing the laser to a smaller spot and spatially selecting the hair follicles in advance may require different laser parameters. Although melanin absorption is ~5 times less at the Nd:YAG wavelength, greater fluence can be delivered to these hair follicles because the rest of the skin is avoided. The 1064 nm wavelength also has greater penetration depths of approximately 1600 μm, so that deeper hair follicles may be destroyed/modified. Although the pulse width of the Q-switched laser

20

is much shorter than the thermal relaxation time of a hair follicle, a thermally-mediated process may not be required to cause permanent hair removal. The mechanism of laser hair removal is not understood, so an adiabatic, mechanical laser-tissue interaction may be just as effective or more effective in permanent hair removal.

The microchip laser is a passively Q-switched, single-mode, diode-pumped, Nd:YAG laser.[12] The Nd:YAG microchip laser array is fabricated with a thin, wide, resonator structure and is pumped by a two-dimensional diode laser array.[13] The diode-laser pump radiation is carried to the microchip via optical fiber. The primary advantage of using the microchip laser is that this source can be fit into a hand-held instrument that comfortably interfaces with the physician and patient. The laser beam has a single spatial mode (TEM$_{00}$) and was focused to a diffraction-limited, 5 μm diameter spot. We have tested two infrared lasers with the following specifications:

Microchip Laser	Wavelength (nm)	Pulse Energy (μJ/pulse)	Fluence (J/cm^2)	Pulse Width (psec)	Peak Power (kW)	Irradiance (GJ/cm^{2}s)
1	1064	120	610	450	267	1360
2	1064	210	1070	700	300	1530

In experiments with mouse skin, the laser converts a dark, absorptive hair into a more highly reflective and/or scattering material. One possible explanation is the hair pigment is photobleached by the high peak powers inherent to these short laser pulses. Another possibility is an increase in surface scatter through modification of the hair surface finish. An example of this effect with from a 1064 nm wavelength laser (120 μJ/pulse, 450 psec pulse width) is shown below. When the same microchip laser beam was focused on a hair follicle, the hair snapped away from the skin leaving little or no surface debris. These observations suggest that the microchip laser may be useful for hair removal.

Another interesting observation was that the 1064 nm beam caused a plasma to form at the laser focus. To better understand whether plasma formation is reasonable for this laser-tissue interaction, we estimate the electric field strength using the following derivation. In general, we calculate the electric field strength in terms of the irradiance of the laser and the tissue material constants. First, the Poynting vector, S, which is analogous to irradiance is given by the following equation:

$$S = \frac{1}{\mu_o} EB \qquad , \qquad (5)$$

where μ_o is the magnetic permeability in free space, $(4\pi \times 10^{-7}\ Ns^2/C^2)$, E is the electric field, and B is the magnetic field. The magnitude of S is the power per unit area crossing a surface whose normal is parallel to S.
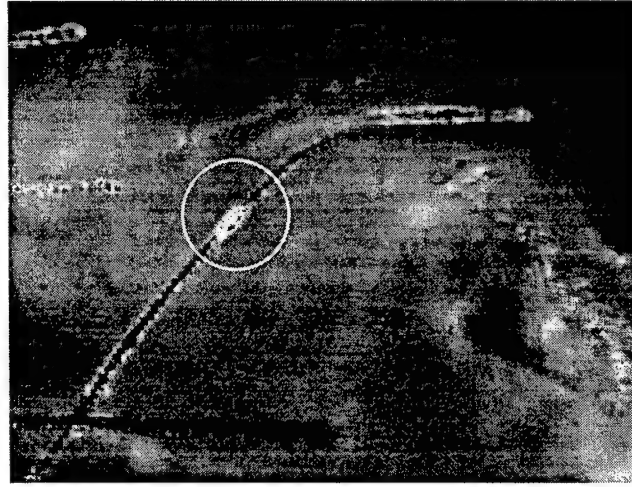

Figure 3: Image of mouse hair modified by 1064 nm microchip laser.

The E and B fields generated by a particle traveling through free space are perpendicular to each other and related by:

$$E = cB \qquad , \qquad (6)$$

where c is the speed of light $(3 \times 10^8\ m/s)$. Using Maxwell's equations, the velocity is given by:

$$c^2 = \frac{1}{\varepsilon_o \mu_o} \qquad , \qquad (7)$$

where ε_o is the electric permittivity in free space, $(8.8542 \times 10^{-12}\ C^2/Nm^2)$, and, as mentioned above, μ_o is the magnetic permeability in free space. Equations 5, 6, and 7 can be manipulated to give the following expression for the Poynting vector:

$$S = \sqrt{\frac{\varepsilon_o}{\mu_o}} E^2 \qquad , \qquad (8)$$

For light propagating through a medium, one must account for the differences in permittivity and permeability in the medium versus these values in free space. These values are related through the absolute index of refraction, n:

$$n = \frac{c}{v} = \sqrt{\frac{\varepsilon\mu}{\varepsilon_o\mu_o}} \qquad , \qquad (9)$$

where ε is the permittivity of the medium, and μ is the magnetic permeability of the medium. For water, which composes 70% of tissue, at 20° C, the refractive index is 1.333. The highest peak power laser had an irradiance of 1.530×10^{16} W/m². Using these numbers, we calculate the field strength as,

$$E = \sqrt{\frac{S\mu_o c}{n}} \qquad , \qquad (10)$$

$$E = 2.08 \times 10^9 \frac{V}{m}$$

In the ablation of brain tissue, the laser tissue interaction with a peak irradiance of 5.7×10^{15} W/m² is reported to be plasma-mediated.[14] Therefore, it is highly probable that this magnitude of electric field strength leads to what we discerned to be plasma breakdown in our experiments.

3. SMART SCALPEL SYSTEM DESIGN

3.1 Prototype system

A diagram of the optical layout for our Smart Scalpel prototype is shown in Figure 4. The desired resolution for the imaging system is 20 μm, ~ 1/10 the diameter of a hair follicle. Two LEDs provide red (660 nm, 10-16 W/sr) and green (565 nm, 0.44-0.63 W/sr) illumination. Each of the two LED outputs is collimated with a convex lens and made colinear with a dichroic beamsplitter. The two beams are then focused and relayed with a biconvex lens to one end of the fiber bundle. Between the lens and the fiber bundle, the light follows a path through a polarizing beamsplitter and a scanning galvanometer. The polarizing beamsplitter is used to pass the light of one polarization for illuminating the tissue and

reflect the orthogonal polarization of light backscattered from the skin to the photodetectors, which makes the underlying hair follicles more apparent. The imaging system galvanometer scans the light onto one end of the fiber bundle.
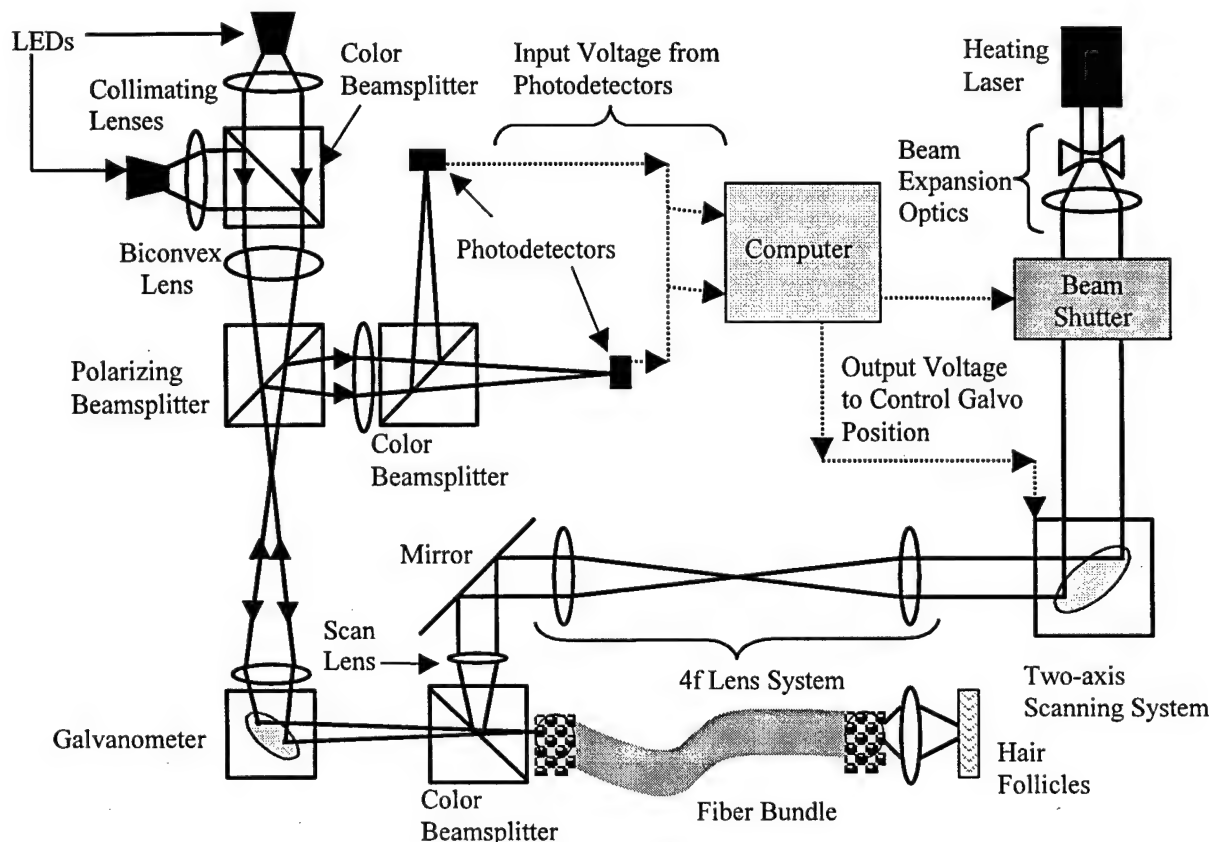

Figure 4: Schematic representation of Smart Scalpel beam scanning and imaging systems.

The light backscattered from the skin is transmitted through the fiber bundle, de-scanned by the galvanometer and reflected by the polarizing beamsplitter. A series of two biconvex lenses transmit the light reflected from the skin surface to two photodectors. A dichroic beamsplitter is used to separate the light into red (melanin-absorbing) and green (hemoglobin-absorbing) wavelengths. Each of these wavelengths is transmitted to a photodetector.

A data acquisition board converts the analog voltage outputs of these two arrays to digital signals, which are used by the computer to identify the spatial locations of hair follicles. The computer runs the necessary image processing algorithms on the array signals to distinguish hair follicles from the rest of the hair shaft and other melanin-rich structures. Once these coordinates are identified, the computer

controls the heating laser beam x-y position via a two-axis galvanometer scanning system. The minimum step response for the current galvanometers is ~ 1 ms, which is comparable to the thermal relaxation time, τ_r. Therefore, during the transit time between targets, a shutter blocks the laser beam to prevent heating of the tissue between hair follicles.
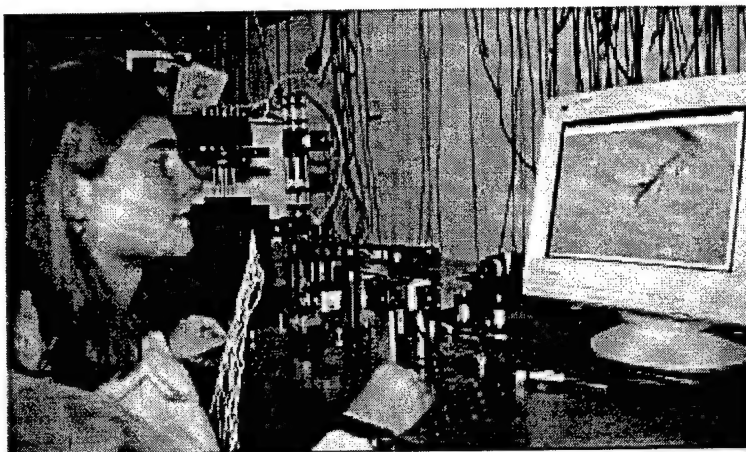

Figure 5: Photograph of prototype Smart Scalpel system.

The heating laser delivery subsystem of the Smart Scalpel is flexible in that any laser can be used with the system. The laser beam is first expanded to the maximum beam diameter that can be scanned by the mirrors on the two-axis galvanometer system. The two galvanometers provide random-access x-y spatial positioning of the laser beam. A 4F lens system is used to transmit the telecentric location between the two galvanometers to the final scan lens, which converts the beam rotation to a displacement scanned on the surface of the fiber bundle. This laser light is then transmitted to the skin surface through the fiber bundle. The final laser spot is approximately 20 μm to provide the desired spatial resolution for the 200 μm hair follicles.

3.2 Control Strategy

Many strategies may be used to deliver this laser energy to the hair follicles. One option is a point detection strategy in which a hand-held instrument is scanned across the skin surface. When a hair follicle is detected, the laser fires at the target. The key requirement for this strategy is that there be minimal time delay between the detection and energy delivery so as the physician scans the surface, the laser beam hits the correct targets. We assume that a physician scans the area at a rate of 10 mm/sec,

and we specify that the laser energy must be delivered within 20 μm ($D_{follicle}/10$) of the detected target. From these requirements, the follicles must be identified and treated within 2 ms.

A second strategy is a stationary device that rests on the skin surface and has a larger field of view than a single hair follicle. With this approach, targets can be identified in advance of the energy delivery. This may be necessary if image processing is required to distinguish the follicle from the rest of the hair. An important consideration for this full field approach is to minimize relative motion between the skin and the Smart Scalpel in the time to image, identify the target coordinates, and steer the laser beam to these locations. The maximum number of targets that can be addressed in one scan can be expressed as:

$$n = \frac{T_{move}}{(T_{exposure} + T_{acquisition} + T_{computer} + T_{galvos} + T_{ill})} \quad , \quad (5)$$

where n is the maximum number of targets per scan, and T_{move} is the period of the highest frequency component that could cause relative motion between the Smart Scalpel and the skin. Time delays in the Smart Scalpel feedback loop include: $T_{exposure}$, the integration time of the line array, $T_{acquisition}$, the time required to acquire and convert the photocurrents to a voltage output, $T_{computer}$, the data acquisition and processing time of the computer, T_{galvos}, the step response of the two-axis scanning system, and T_{ill}, the time required for the laser to thermally or mechanically treat the hair follicles. The highest frequency movement that has been identified is tremor. Tremor is classically said to be a 10 Hz quasi-sinusoidal displacement, although the frequency of tremor varies among different body parts and different people.[9] This movement requires that the region of skin be scanned and treated within ~100 ms.

3.3 Miniaturization

Once the prototype system is tested and the optimal control strategy is determined, the Smart Scalpel will be miniaturized to interface more comfortably with the physician and patient. If a point-detection strategy is used, our design can be implemented as a hand-held surgical instrument, which the physician scans over the skin surface (Figure 6). Two light-emitting diodes (LEDs) at desired wavelengths (i.e. 565 nm and 650 nm) and one or two photodetectors are mounted within the instrument. If LEDs are used to illuminate the skin, we could turn each LED on and off 180 degrees out of phase. With this method, a first reading from the photodetector would be acquired for a wavelength of high melanin absorption. A second photocurrent measurement would then be taken with the skin illuminated at the

565 nm wavelength to normalize the absorption measurement. We could utilize analog and/or digital circuitry to compare the photodetector currents and determine whether the instrument is located above a hair follicle or surrounding tissue. We must investigate the feasibility of this approach by determining the maximum time delay between initialization of the detection procedure and delivery of laser energy. If we want to direct the laser beam within 20 μm of the detected target, this delay cannot be longer than 2 msec, assuming a scan rate of 10 mm/sec.
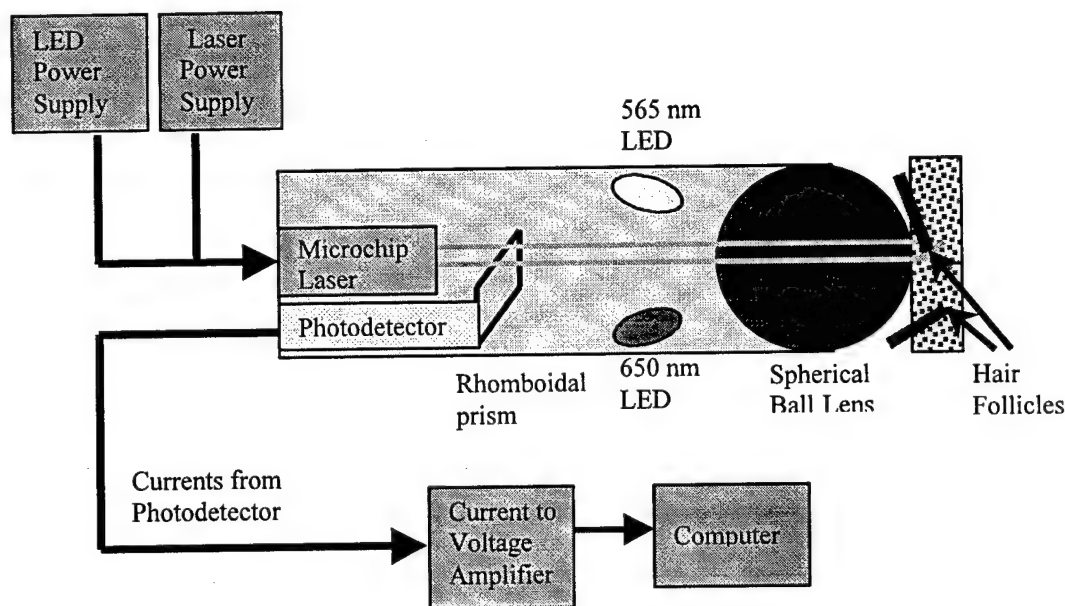
Figure 6: Schematic of miniaturized Smart Scalpel.

As hair follicles are detected, laser energy is delivered to the instrument via a single mode fiber optic. If the microchip laser is used, the laser can be mounted within the instrument. A ball lens contacts the skin surface and focuses the laser light to the hair follicle. The ball lens also collects the reflected light, which is directed to the photodetectors.

4. FUTURE WORK

Project deliverables for the Smart Scalpel system focus on characterizing the spectroscopic sensitivity of the system and then testing a wide range of control strategies, lasers, and scan areas to determine the optimal treatment parameters.

- The spectroscopic sensitivity of the prototype system will first be evaluated with test patterns of hair. The optical reflectance measurements, the computer will identify the spatial coordinates of the hair and control the laser scanning system to cover only the identified targets.

- When the system can accurately recognize the hair and scan the laser over only these coordinates, the next step will be to test the imaging and treatment components of the system using an animal model. Through these experiments, much will be learned about the laser wavelength, fluence, pulse width, control strategy, and other parameters leading to the optimal therapeutic outcome.

- The final deliverable for the hair removal project is to develop a robust system that interfaces well with both physician and patient in a clinical setting.

REFERENCES

1. R.N. Richards, U. Marguerite, and G. Meharg, "Temporary Hair Removal in Patients with Hirsutism: A Clinical Study," *Cutis*, 45, pp. 199-202, 1990.

2. R.F. Wagner, "Physical Methods for the Management of Hirsutism," *Cutis*, 45, pp. 319-26, 1990.

3. R.R Anderson, "Laser Medicine in Dermatology," *Journal of Dermatology*, 23(11), pp. 778-782, November, 1996.

4. J.B. Dawson, D.J. Barker, D.J. Ellis, E. Grassam, J.A. Cotterill, G.W. Fisher, and J.W. Feather, "A Theoretical and Experimental Study of Light Absorption and Scattering by *in vivo* Skin," *Phys. Med. Biol.*, 25(4) pp. 695-709, 1980.

5. R.R. Anderson, and J.A. Parrish, "The Optics of Human Skin," *The Journal of Investigative Dermatology*, 77, pp. 13-9, 1981.

6. M.C. Grossman, C. Dierickx, W. Farinelli, T. Flotte, and R.R. Anderson, "Damage to Hair Follicles by Normal-Mode Ruby Laser Pulses," *Journal of the American Academy of Dermatology*, 35, pp.889-894, 1996.

7. C. Dierickx, M.C. Grossman, W. Farinelli, and R.R. Anderson, "Permanent Hair Removal by Normal-Mode Ruby Laser," *Arch. Dermatol.*, 134, pp. 837-842, 1998.

8. C.A. Nanni, T.S. Alster. "Optimizing treatment parameters for hair removal using a topical carbon-based solution and 1064-nm Q-switched neodymium:YAG laser energy," *Arch. Dermatol.*, 133(12), pp. 1546-9, December 1997.

9. K.P. Watts, R.G. Fairchild, D.N. Slatkin, D. Greenberg, S. Packer, H.L. Atkins, and S.J. Hannon. "Melanin Content of Hamster Tissues, Human Tissues, and Various Melanomas," *Cancer Research*, 41(2), pp. 467-472, February 1981.

10. A.L. McKenzie, "Physics of Thermal Processes in Laser-Tissue Interactions." *Phys. Med. Biol.*, 35(9), pp. 1175-1209, 1990.

11. W.F. Cheong, S.A. Prahl, and A.J. Welch, "A Review of the Optical Properties of Biological Tissues," *IEEE Journal of Quantum Electronics*, 26(12), pp. 2166-2185, 1990.

12. R.S. Afzal, A.W. Yu, J.J. Zayhowski, and T.Y. Fan, "Single Mode, High Peak Power, Passively Q-Switched Diode-Pumped Nd:YAG Laser," *Optics Letters*, 22(17), pp. 1314-1316, 1997.

13. C.D. Nabors, J.J. Zayhowski, R.L. Aggarwal, J.R. Ochoa, J.L. Daneu, and A. Mooradian, "High-Power Nd:YAG Microchip Laser Arrays." *Optical Society of America Proceedings on Advanced Solid-State Lasers,* 13, Proceedings of the Topical Meeting, pp. xvii+391, 234-6, 1992.

14. J.P. Fischer, J. Dams, M.H. Gotz, E. Kerker, F.H. Loesel, C.J. Messer, M.H. Niemz, N. Suhm, J.F. Bille. "Plasma-Mediated Ablation of Brain Tissue with Picosecond Laser Pulses." *Applied Physics B.* 58, pp. 493-499, 1994.

Patient Monitoring and Diagnosis

CHAPTER 6

Noninvasive Blood Glucose Analysis Using Near Infrared Spectroscopy
K. Youcef-Toumi, V. Saptari

Noninvasive Blood Glucose Analysis using Near Infrared Spectroscopy

Prof. Kamal Youcef-Toumi
Principal Investigator

Vidi A. Saptari
Graduate Research Assistant

I. Introduction

This work is motivated by the profound need of a noninvasive way to measure glucose in the blood. A noninvasive glucose monitoring device would provide a safer and a more convenient method to treat and control diabetes. The goal of diabetes therapy, within and outside hospital, is to approximate the 24-hour blood glucose profile of a normal individual, which necessitates continuous monitoring.

Several optical techniques have received considerable attention over several years, which include Raman, absorption, scattering and polarimetry spectroscopy. The most attractive feature of optical-based sensors is the inherent non/minimal invasiveness. Furthermore, unlike chemical-based sensors, they do not require reagents, and therefore simplify measurements.

In this report, overview of our system design and characterization is included, as well as results from preliminary experiments indicating fundamental feasibility of the technique. Finally, we discuss the obstacles and issues associated with this work that need to be addressed.

II. System Description

In general, spectroscopic techniques measure interactions between lights and matters. They differ between one and another in the types of interactions they measure. Initially, we studied and considered the various spectroscopic methods to determine the most suitable system for this project. Overview of each technique has been presented in the previous report (phase II-1st year).

Raman and absorption spectroscopy were two of the most promising techniques to be used as noninvasive glucose sensors. Raman technique is advantageous with respect to the specificity it can provide, making it a better technique for mulicomponent blood analysis. However, the extremely weak signals causes it to be inappropriate for *in vivo* measurements, as this means that to get an acceptable signal-to-noise ratio, the power of the light directed onto the tissue has to be unsuitably high. Since glucose is our sole concern in this project, the specificity that Raman technique provide may not be very important. Therefore, the absorption technique was selected for this work.

(i) Fundamental Aspect of Absorption Spectroscopy

The underlying principle of this method is that when an electromagnetic radiation such as light interacts with a sample, its frequency components (spectra) are altered. This is due to the fact that each molecule absorbs light at specific frequencies corresponding to

their vibrational and rotational oscillations. This property acts as a fingerprint, which identifies the molecular composition of the matter. Furthermore, the concentration of a particular molecule has a linear relationship with the intensity of the absorbance peaks (Beer's law), which can be used to predict the concentration of certain molecules in the substance [1]. It states that the intensity of an absorption peak is given by:

$$A = \varepsilon(\lambda)lc$$

where ε is the absorptivity coefficient, l is the path length and c is the concentration.

(ii) System Characterization

- **Wavelength Source Consideration**

Since each molecule absorbs light at definite frequencies, the first task would be to determine the wavelength range that is appropriate. Glucose has many absorption peaks from the near infrared up to the mid-infrared ($800\,nm$ to $10\,\mu m$), although not all of them are specific for only this molecule. The peaks in the mid-IR region correspond to the fundamental vibrations, whereas in the near-IR, they correspond to overtone and combination vibrations. This results in weaker and broader peaks in the later case.

However, the wavelength region of the near infrared possesses some properties that make it more suitable for non-invasive *in vivo* diagnosis as compared to the mid-infrared. In this range, water, which is the dominant component of blood and tissue absorbs mid-infrared light strongly, allowing only very shallow penetration depths [2]. Therefore, mid-infrared spectroscopy has only been considered as an invasive technique up to date. Near infrared light, on the other hand, can penetrate into the skin up to 1 cm. Looking at results from other investigators [2-4], and also our own results from a preliminary experiment (see Part V of this report), an appropriate wavelength range to consider is from about 700 *nm* to 1900 *nm*.

- **Fourier-transform Spectrometer vs. Dispersive Spectrometer**

A couple of ways of decomposing a light into its spectrum are dispersive and Fourier-transform interferometric techniques. Dispersive spectrometers use gratings or prisms to disperse the light into a spectrum of its component wavelengths. A slit is then used to select which narrow "slice" of them is allowed to strike the detector. An optical diagram is shown in figure 1.
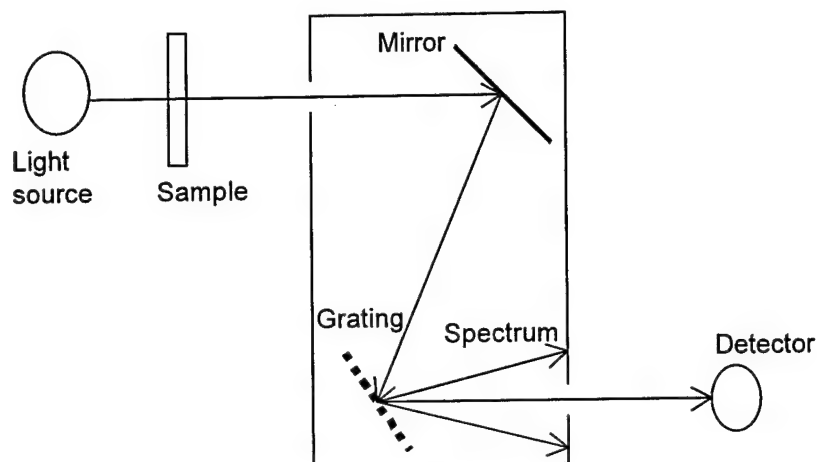
Figure 1. Dispersive spectrometer

An alternative way to decompose the light into its wavelength constituents is to use an interferometric technique. A schematic of a typical spectrometer utilizing Michelson interferometers is shown in figure2.
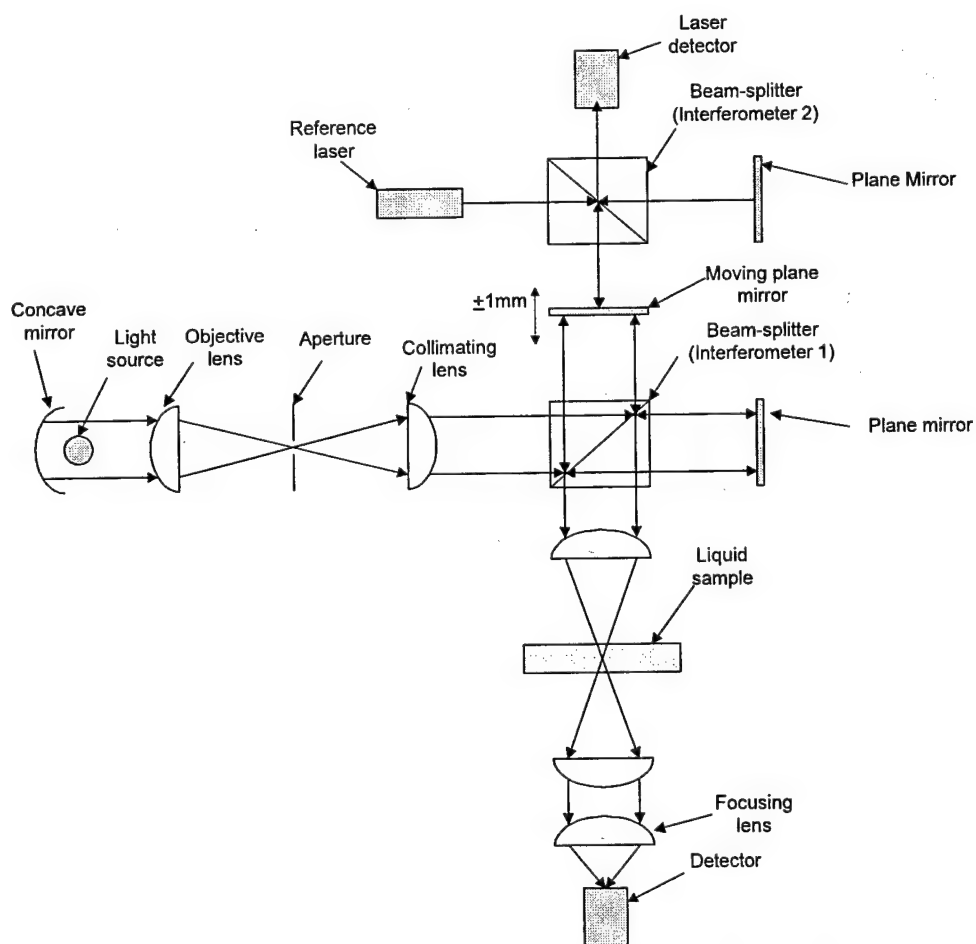

Figure 2. A schematic of a typical Fourier transform spectrometer

3

A collimated beam from a source enters the interferometer and onto the beamsplitter, at which half is reflected and half is transmitted. The beams reflected off of the moving and fixed mirror are then recombined on the beam splitter, and imaged onto the sample and finally onto the detector. As one of the mirror moves, the two recombined beams undergo amplitude interference due to the path difference. This produces interferogram, which is seen by the detector and recorded by a computer. This interferogram is then Fourier-transformed to give the spectrum of the light transmitted out of the sample. The second interferometer with the laser and the laser detector is used as position and velocity control of the moving mirror.

In a dispersive system incorporating a grating and an exit slit, only one spectral element is sampled by the detector at a time. In contrast, in Fourier transform spectroscopy, one examines all wavelengths arriving at the detector simultaneously, which results in higher throughputs. Another advantage of interferometric technique is the superiority of its signal-to-noise ratio. In theory, all spectrometers can show an improved signal-to-noise ratio if spectra are averaged. However, this relies on the fact that the spectra can be exactly superimposed. Any displacement error between spectra can cause band shapes to be distorted, and as a result, the signal-to-noise ratio will fail to improve. Scanning monochromators are subject to mechanical wear and jitters, which may cause significant displacement errors. Interferometric systems utilizing He-Ne laser for data acquisition trigger and motion control are very stable. Data can be acquired at very precise path differences, even if there are some jitters in the velocity of the moving mirror. It is for these reasons that we choose Fourier transform spectrometer for our application.

III. Near Infrared Fourier Transform Spectrometer Design Overview
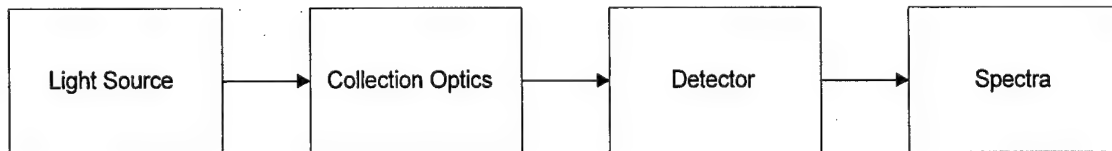

Figure 3. Functional blocks of Fourier transform spectrometer

In this part, calculation of design parameters of the Fourier transform spectrometer is presented. Parameters are calculated with the aim to maximize spectrometer signal-to-noise ratio, which is given by [5]:

$$SNR = \frac{U_{\bar{v}}(T).\Theta.\Delta\bar{v}.t^{1/2}.\zeta}{NEP} \tag{1}$$

where,

$U_{\bar{v}}(T)$ is spectral density at wavenumber $\bar{v}$ from a black body source at a temperature T $(W / sr.cm^2.cm^{-1})$

Θ is the throughput of the system $(cm^2.sr)$

4

$\Delta \bar{v}$ is the resolution of the spectrum (cm^{-1})

t is time in seconds

ζ is the efficiency accounting for losses due to the optical components

NEP stands for noise equivalent power, which is a sensitivity *figure-of-merit* of the detector $(W.Hz^{-1/2})$

To begin the design analysis, we start by looking at the two constrains or fixed parameters, namely:

- the resolution required and
- the wavenumber range to work in

For the near infrared absorption measurements, resolution of 32 cm^{-1} is usually sufficient [2] since the peaks are broad. For this design, the best resolution achievable is decided to be 10 cm^{-1}, and from before, the wavenumber range is between 14300 cm^{-1} and 5000 cm^{-1}, which corresponds to wavelength range between 700 and 2000 nm.[a]

1. Detector Consideration

The most basic and unavoidable of all types of noise in a spectrum measured using a Fourier transform spectrometer is detector noise. The sensitivity of infrared detectors is commonly expressed in terms of its noise equivalent power (NEP). It shows the incident power required to generate a response equal to the noise level of the detector system (to give SNR equal to one) as measured at the amplifier output at a given frequency. It is expressed as the noise current in units of $(A/\sqrt{Hz})$ divided by the responsivity in (A/W), with the resulting units of $(W.Hz^{-1/2})$. A better detector has a smaller NEP number since this means that it gives a higher SNR for a certain incident power.

Another detector parameter that needs to be considered is its area. Larger area means more light can be collected in a given time. However, it also increases the NEP.

2. Interferometer and Collection Optics Consideration

Given the resolution required, one can calculate the path difference l, which in the case of a double-sided scanning interferometer, is equal to the total distance of the moving mirror. Resolution is inversely proportional to the distance moved [5].

$$l = \Delta \bar{v}^{-1} \qquad (2)$$

Due to imperfect-collimated beam reaching the beam splitter and the mirrors, wavefront-splitting interference occurs, which results in circular fringes when the mirrors and the beam splitter are aligned. Unless we limit the detector to receive only the central fringe, the data would be useless. To avoid this, we use an aperture as a spatial filter. The maximum aperture diameter allowed is given by [6]:

[a] $Wavelength(nm) = \dfrac{1}{Wavenumber(cm^{-1})} * 10^7$

$$D_A = 2\sqrt{\frac{f^2}{l.\bar{v}}} \qquad (3)$$

where f is the focal length of the lens which collects and collimates the light entering the interferometer.

Before we decide on the focal length, we consider the detector throughput. Throughput is defined as:

$$\Theta_D = A_D \Omega_D \qquad (4)$$

where Ω_D is the solid angle of the beam being focused on the detector, with a unit of steradian (sr). For a focusing lens with a diameter d and a focal length f:
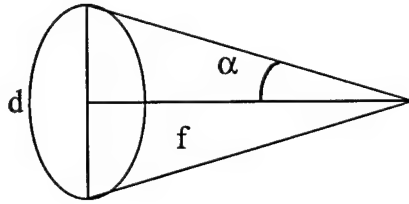
$$\Omega = 2\pi\alpha^2 \qquad (5)$$

Detector area should be chosen so as to optimize both NEP and detector throughput.

The rest of the spectrometer is designed such that the throughput is limited by the detector throughput. Spectrometer throughput is defined as[5]:

$$\Theta_S = \frac{2\pi A_B \Delta\bar{v}}{\bar{v}_{max}} \quad (cm^2 sr) \qquad (6)$$

where A_B is the area of the collimated beam entering the interferometer, and $\bar{v}_{max}$ is the highest light wave frequency in the spectrum. Equating (4) and (6) gives the relationship between α and A_B. α should be chosen so as to maximize the collection power with the limit being the optical aberrations. Going back to equation (3) we can now calculate the aperture diameter.

IV. Overview of Hardware Implementation

The system is built on a 0.9 X 0.6 m optical breadboard with modular opto-mechanical components as structure. Light source is a 100 Watts tungsten-halogen, which contains wavelengths from the visible up to the near infrared. Before entering the beam splitter, the beam is focused onto a variable-diameter aperture and collimated by a pair of plano-convex calcium fluoride lenses with focal lengths of 150 mm (Coherent-Ealing).

The interferometer is constructed from two plane surface mirror with silver coating and a broad-band nonpolarizing cube beam splitter (OptoSigma). One mirror is fixed on a micrometer-driven translation stage, and the other is mounted on a linear guide, driven by a voice-coil actuator (BEI). Position and velocity control is accomplished by a second interferometer using HeNe laser as reference.

The beam output from the interferometer is focused onto the sample and finally onto the detector. Detector used is a 1mm diameter, thermoelectrically cooled InGaAs (EG&G Judson), with an operating wavelength range between 800 and 1900 nm. Cooled InGaAs detectors give lower noise and hence smaller NEP.

V. Preliminary Experiment

A preliminary experiment was performed using a commercial infrared spectrometer. The aims of the experiment were to determine the appropriate wavelength range to work in, as well as for us to confirm that this technique was fundamentally feasible. However, no quantitative studies/predictions were performed. Changes in the spectra due to the addition of glucose were simply recorded and inspected by eye.

(i) Material and Method

A Nicolet Magna-IR 860 equipped with a DTGS detector and a CaF_2 beamsplitter was used to determine the effect of glucose on water and human blood serum near infrared spectra. Spectra were obtained in transmission mode, with the resultant transmission spectra converted to an absorbance plot. Sample was placed in a 10 mm pathlength quartz cuvette. For each solution, spectra of the pure solution were initially recorded. A known composition of D-glucose powder was then added.
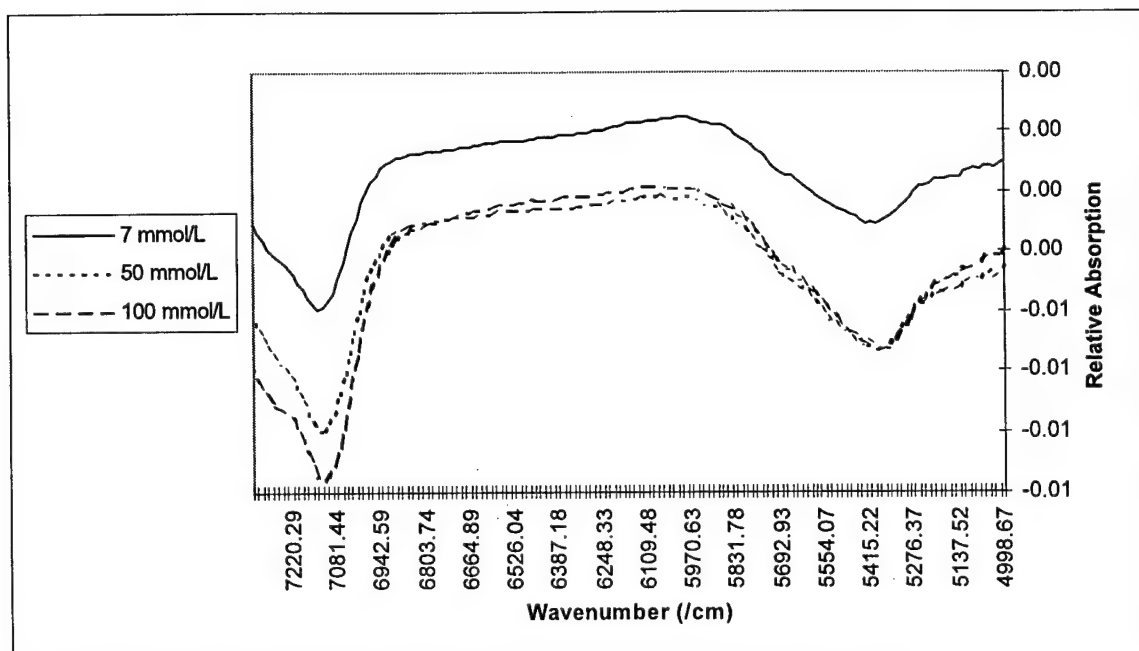
(ii) Results

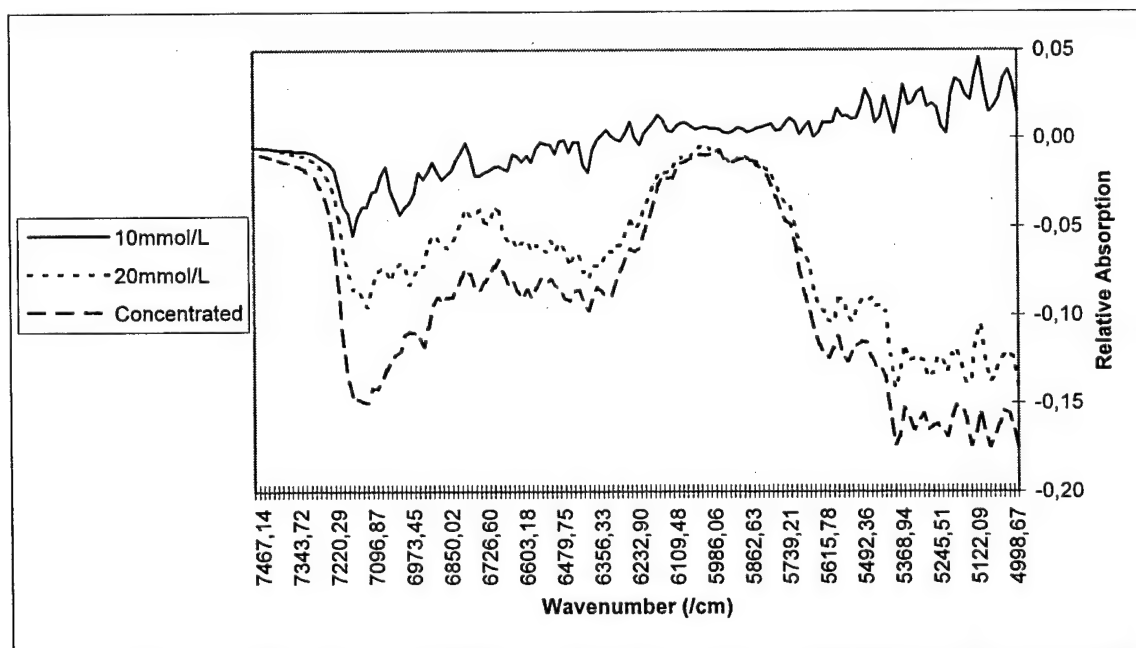Figure 4. Differential spectra of water solution due addition of glucose

Figure 5. Blood serum differential spectra due to addition of glucose

Figure 4 and 5 represent *glucose-with-water* spectra and *glucose-with-serum* spectra respectively. In each graph, spectra of the solvent was subtracted. Using this spectrometer, changes in the spectra due to glucose could be recorded down to 7 *mmol/L* in the case of water solution, and 10 *mmol/L* for the blood serum, which corresponds to the higher physiological range. It can be seen that the presence of glucose clearly affects

the absorption at the wavenumber around 7150 cm^{-1} (1400 nm). It also seems to have some influence at around 5400 cm^{-1} (1850 nm), which should be further investigated.

These results also prove that near infrared absorption spectroscopy is a feasible technique for glucose measurements. However, there are still significant problems that need to be addressed before realization of an accurate monitoring device is possible. The next section discusses the obstacles and our plans in this research.

VI. Main Obstacles

- Low accuracy due to low signal-to-noise ratio:
  For physiologic concentration, glucose produces little quantifiable change in the spectral output in the near infrared region, due to the fact that glucose concentration level in the blood is very small, about 0.1% for normal individuals. Results from several groups [3, 4, 7] show that the required accuracy for safe blood glucose measurements has not been achieved.

- Erroneous readings due to tissue variations:
  Results have shown that tissue variations can cause significant changes in the output spectra, resulting in the irreproducible data and the need for frequent calibration [3, 4, 7]. These changes may include skin temperature changes, hemoglobin concentration, contact pressure of the measuring probe, etc.

VII. Research Plan

To deal with the first obstacle, one needs to start with a reliable instrument that has high sensitivity or signal-to-noise ratio. Although many commercial infrared spectrometers, such as those used by other researchers give good performance in the mid-infrared region, they usually are not intended for near infrared measurements. In addition, most of the commercial spectrometers do not give the flexibility required for research purposes. It is for these reasons that we decided to build our own spectrometer. First, its performance characteristics would be optimized for our work. Second, since the instrument is made up of modular components, it would be flexible enough for almost any changes in configuration. For example, the sample can be positioned anywhere, and addition of other component such as a confocal microscope can easily be implemented.

The next task would then be to deal with the problem of irreproducibility of data due to tissue variations. Since blood vessels lie underneath the skin and tissue, the back-scattered light will also contain skin and tissue information. A highly reliable system will require a robust discrimination strategy which can distinguish the useful information from the blood, and disregard that from the skin and tissue. Upon completion of the instrument development, we would attempt to address this issue, which may require design of new hardware and/or procedural scheme.

References

1. Smith B., "Fourier Transform Infrared spectroscopy", CRC Press, Florida, 1996
2. Marbach R. et. al., "Noninvasive Blood Glucose Assay by Near Infrared Diffuse Reflactance Spectroscopy of the Human Inner Lip", Appl. Spect. 47 (7): 875-881, 1993
3. Ham F.M. et. al., "Multivariate Determination of Glucose Concentrations from Optimally Filtered Frequency-warped NIR Spectra of Human Blood Serum", Physiol. Meas. 17: 1-20, 1996
4. Robinson M.R. et. al., "Noninvasive Glucose Monitoring in Diabetic Patients: A Preliminary Evaluation", Clin. Chem. 38 (9): 1618-1622, 1992
5. Griffiths P.R., de Haseth J.A., "Fourier Transform Infrared Spectrometry", John Wiley & Sons, 1986
6. Thorne A.P., "Spectrophysics", Chapman & Hall, New York, 1992
7. Muller U.A. et. al., "Non-invasive Blood Glucose Monitoring by Means of Near Infrared Spectroscopy: Methods for Improving the Reliability of the Calibration Models", Artif. Pancreas and Related Tech. In Diabetes and Endocrinology 20 (5): 285-290, 1997

Home Treatment

CHAPTER 7

Glucose Sensor and Insulin Delivery Device
T. Kanigan, C. Brenan, I. Hunter

GLUCOSE SENSOR AND INSULIN DELIVERY DEVICE

PROGRESS REPORT FOR YEAR 1

Tanya Kanigan, Colin Brenan and Ian Hunter

INTRODUCTION

OBJECTIVES

Our goal is to develop a technology for continuous monitoring and control of blood glucose levels, which is minimally invasive and amenable to mass-production and miniaturization. The device must be able to accurately measure glucose concentration and to dispense therapeutic amounts of insulin using the glucose level measurements as feedback.

COMMERCIAL TECHNOLOGY FOR DIABETES SELF-MONITORING AND TREATMENT

Diabetes afflicts approximately 120 million people worldwide and approximately 16 million people within the United States. The American Diabetes Association recommends that diabetics maintain glucose levels of between 80 to 120 mg/dl (4.4 to 6.7 mmol/L) before meals and 100 to 140 mg/dl (5.6 to 7.8 mmol/L) before bed. High glucose levels are usually treated with a combination of diet, exercise and oral medication; insulin injections are used in only about 20 % of cases.

Roughly 65 % of American diabetics routinely measure their blood glucose levels, on average once per day, using personal glucose monitors. The measurement is made by lancing one's fingertip to extract a drop of blood (typically 5-10 µl) and applying it to a test strip coated with glucose oxidase and other reagents. The test strip is then inserted into a portable meter, which displays the glucose concentration in mg/dl.

1

Glucose oxidase is an enzyme that catalyzes the oxidation of glucose to gluconic acid and hydrogen peroxide. The hydrogen peroxide may then be assayed by reacting it with other chemicals to produce a colored chemical species, the concentration of which, in turn, is measured photometrically. Alternatively, the hydrogen peroxide may be detected electrochemically (amperometrically). Most commercially available glucose monitors use one of these two methods.

Several companies are currently developing minimally invasive glucose monitors based upon either vibrational spectroscopy of tissues and/or the harvesting of interstitial fluid (see Table 1). Cygnus, for example, is currently seeking FDA approval for its Glucowatch, a device that extracts glucose through the skin using a small electric field and then detects it electrochemically. Implantable insulin pumps are commercially available from several sources. These systems rely on the user to adjust the insulin dose to counteract blood glucose fluctuations from meals. No device yet combines both glucose monitoring and insulin delivery to achieve a closed-loop system (i.e., an artificial pancreas), even though such a device is the "Holy Grail" of diabetes treatment.

Table 1 Commercial non-invasive and minimally invasive glucose monitoring systems under development (IF = intersitial fluid).

Company	Product	Website	Operating Principle	FDA Status
Biocontrol Technology	Diasensor 1000	www.bico.com	Fiber-optic based IR reflection	Under FDA review
Cygnus	Glucowatch	www.cygn.com	Reverse iontophoresis of IF	Plans to submit in 1998
Pacific Biometrics	SalivaSac	www.pacbio.com	Detects glucose in saliva	Not yet submitted
SpectRx	SpectRx	www.spectrx.com	Extracts IF	At least 2 years before submission
Integ	LifeGuide	www.integonline.com	Measures glucose in IF using Far IR	Plans to submit in 1998

DETERMINING GLUCOSE LEVELS FROM THE VIBRATIONAL SPECTRA OF TISSUES

Vibrational spectroscopic techniques such as near IR and Raman spectroscopy use intense radiation sources to probe the vibrational modes of molecules. The vibrational spectra of organic molecules such as glucose are complex and unique; however, spectra from molecules with similar structures are similar enough that a single vibrational band cannot be used to distinguish between them. In both blood and in interstitial fluid, glucose presents a weak signal upon a large temperature-sensitive background of water and other biological chemicals such as urea, cholesterol and alcohols (Cote, 1997). As such, quantifying the relative contribution from glucose is challenging. Even when multivariate spectral analysis techniques are used, mean square prediction errors of 13.2 mg/dl at best have been reported for *in vitro* blood serum measurements (Ham, 1996).

In the case of near IR spectroscopy small variations in sample temperature, such as those that can be induced by the radiation source, cause frequency shifts in the water bands that dominate the spectra (Wang, 1998). A temperature variation of less than a degree cause a more significant absorption changes in the relevant spectral region than do 20 mg/dl variations in glucose concentration. Minimizing this effect with careful temperature control is feasible, but will also complicate device design and increase measurement time.

Raman spectroscopy is much less sensitive to the presence of water and is typically preferred over IR spectroscopy for aqueous-based samples. Raman spectroscopy measures inelastic scattering from induced dipole fluctuations that occur when molecules vibrate. As this is a much weaker effect than IR absorption, a visible or near IR laser is used as the excitation source. Advances in solid state laser and detector technologies are making Raman spectrometers smaller and more affordable. For example, a Raman spectrometer can be purchased from Ocean Optics of Dunedin, FL for $10K – an appropriate laser excitation source costs an additional $10-20K. Although such a system

may soon be appropriate for use in the home or clinic, this technology is not suitable for wearable technology.
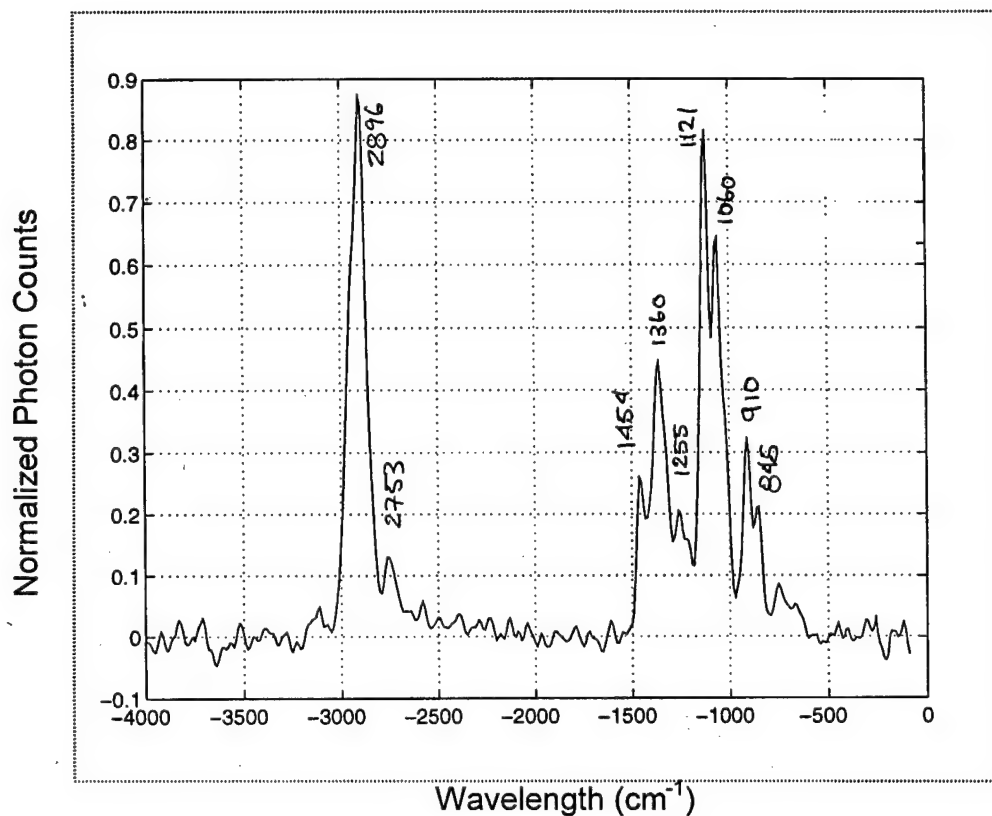

Figure 1 Raman spectrum of 2.48 M glucose in water (water background subtracted). Spectrum was collected with laser power of 31 mW and an integration time of 50 ms. Note that this spectrum was collected with a glucose concentration approximately 1000 times more concentrated than typical blood levels.

Recently, a strong correlation has been established (Bantle, 1997) between blood glucose levels and glucose levels in interstitial fluid withdrawn from dermal tissue. Glucose levels in the interstitial fluid lag values in capillary blood by approximately 15 minutes. Since this fluid may be extracted from the dermis below the stratum corneum, but above blood vessels and nerves, analysis of this fluid provides a bloodless, painless approach to measuring physiological glucose levels. Similarly, drugs can be introduced into the blood stream by injecting them just below the protective stratum corneum. The drugs diffuse through the dermal tissues and into blood vessels.
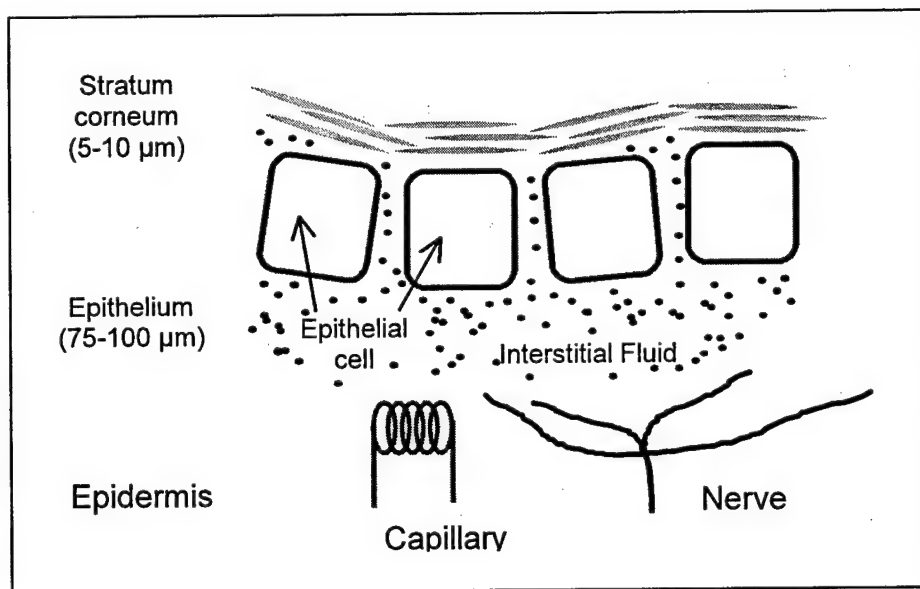

Figure 2 Simplified model of the skin.

DEVICE DESIGN

Our goal is to produce a glucose monitoring/insulin delivery device which can be worn continuously by the diabetic patient. The system will attempt to mimic the glucose control function of the pancreas by continously adjusting insulin levels to maintain appropriate glucose concentrations in the blood in response to the feedback from the glucose sensor. The same unit can function as a tool for doing system identification: by recording how a patients glucose level responds to injections of insulin, the impulse response function to an insulin bolus can be determined for that individual. This information may be useful both for diagnosing diabetes and for tailoring the insulin therapy regimen.
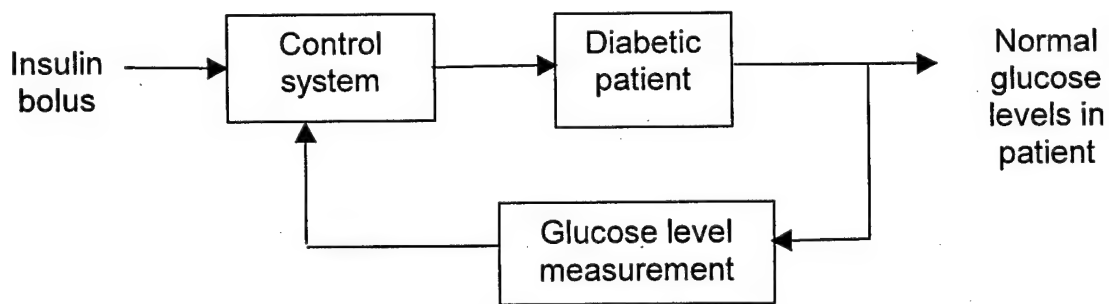

Figure 3 Control feedback loop for the glucose monitoring/insulin delivery device.

More specifically the device will consist of a microneedle array that penetrates the stratum corneum and functions both as an insulin delivery conduit and as a glucose sensor, an insulin reservoir, a valve for controlling the flow of insulin, a microprocessor and necessary electronic hardware. If diffusion alone provides sufficient flow rates for insulin therapy, the injection of insulin will be controlled by a microvalve alone. If not, a miniature pump will be added to the device. Our initial design (shown in Figure 3) was based upon an array with two sets of needles, one for withdrawing interstitial fluid into

6

a glucose sensing region, and another for injecting small doses of insulin into the epidermis. More recently, we learned of amperometric glucose sensors constructed from stainless steel hypodermic needles capable of measuring glucose concentrations in a physiologically relevant range (0 – 400 mg/dl) (Yang, 1998). We are now hoping to use the exterior of the microneedle array as selective glucose electrodes. If this proves feasible, the microneedle array interior will be used to deliver insulin exclusively.
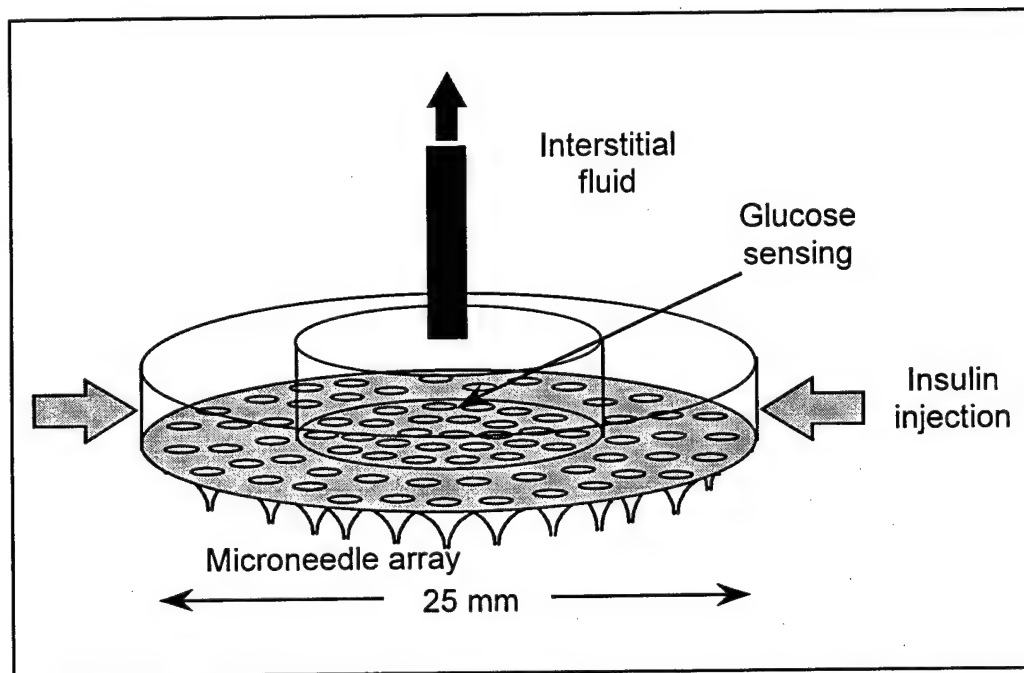

Figure 4 Schematic drawing of an initial design for the glucose sensor/insulin delivery device.

RESEARCH PLAN AND APPROACH

In the second year of this project we plan to fabricate two separate prototype microneedle devices, one which measures glucose levels in interstitial fluid, another which both measures glucose levels and delivers insulin through a computer-controlled microvalve. In order to accomplish this goal several design issues surrounding key technologies must be addressed.

INSULIN THERAPY REGIMEN

The following questions must be answered:

- How much insulin is typically delivered and at what intervals?

- How fast do blood glucose levels change in response to insulin injections?

- What factors affect the efficacy of insulin (i.e., temperature, adsorption onto surfaces, light, etc.)

GLUCOSE SENSOR DESIGN

A suitable glucose detection method must be found which is fast (<30 s measurement time), accurate (<5% variation from a reference method), non-toxic and miniaturizable. Biosensors which combine enzymatic and electrochemical (amperometric) detection can measure glucose at physiological levels (0-400 mg/dl) in 10 s with a sensitivity of 40 nA/mM (Yang, 1998). Since such sensors can be produced from thin films of polymer and polymer containing glucose oxidase on a conducting surface, they are inherently suitable for minitiarization.

SKIN TISSUE PROPERTIES

The mechanical impedance of the upper epidermal layers to the flow of interstitial fluid must be ascertained. This will determine whether active pumping must be applied

to withdraw interstitial fluid out of the tissue (should this be necessary for glucose detection) and whether pressure is necessary to force insulin into the tissue. A rough estimate may be obtained from previous reports: 200-400 mmHg of differential pressure will produce an interstitial fluid flow of 0.2-0.3 μL min^{-1} cm^{-1} (Svedman, 1998). Further experimentation will be necessary to refine this figure.

INSULIN INJECTION SYSTEM

Microvalve technologies (such as conducting polymer, shape memory alloy) and pumping mechanisms (if necessary) must be evaluated for applicability.

MICRONEEDLE ARRAY DESIGN AND MANUFACTURE

The footprint of the microneedle array will be based upon the rate at which fluid must be delivered during insulin therapy and the dimensions of individual microneedles. Needles must be both sharp enough and strong enough to penetrate the stratum corneum.

It highly desirable to keep the manufacturing costs for the microneedle array down such that it can be treated as a disposable. This will help minimize degradation of insulin, glucose oxidase and other array constituents. A survey of various manufacturing techniques that could be used to fabricate the microneedle arrays is presented in Table 2.

Table 2 Potential manufacturing techniques for producing microneedle arrays.

Fab. Technique	Material
LIGA machining	Almost any
Excimer laser machining (MIT)	Almost any (eg glass, plastic)
Si lithography (deep etch) (Bosch process)	Si wafers
Differential etching (MIT/Schott)	Glass
PCB drilling (MIT/Promacad)	Au/Cu clad glass epoxy (or teflon,..) (round only)
Polymer milling (MIT)	any machinable plastic (eg Delrin) (round only)
MicroEDM wire & sink (MIT)	Any conductor (eg stainless steel)
MicroEDM wire & ram (MIT)	any thermal plastic (eg polycarbonate)

SUMMARY

We have shifted the focus of our work in the area of glucose monitoring away from the development of a non-invasive spectroscopic tissue probe suitable for use in the home or clinic. Instead we are concentrating on the development of an "artificial pancreas," a minimally invasive device for glucose measurement and closed-loop insulin delivery. The device will be capable of modulating the flow of insulin through a patient's skin while simultaneously monitoring physiological glucose levels in interstitial fluid. This approach will permit system identification and control techniques to be used to continuously maintain glucose levels at an optimum level.

REFERENCES

Bantle, John P., and Thomas, William (1997), "Glucose measurement in patients with diabetes mellitus with dermal interstitial fluid." *J. Lab. Clin. Med.,* **130**, 436-41.

Cote, Gerard L. (1997), "Noninvasive optical glucose sensing – an overview," *Journal of Clinical Engineering,* **22** 253-259.

Ham, Frederic M., Cohen, Glenn M., Kostanic, Ivica and Gooch, Brent R. (1996), "Multivariate determination of glucose concentrations from optimally filtered frequency-warped NIR spectra of human blood serum." *Physiol. Meas.,* **17**, 1-20.

Svedman, Paul and Svedman, Christer (1998), "Skin Mini-Erosion Sampling Technique: Feasibility Study with Regard to Serial Glucose Measurement." *Pharmceutical Research,* **15**, 883-888.

Wang, Quian, Tague Jr., Thomas T., and Melling, Peter (1998) "Determination of glucose concentrations at physiological levels by FT-NIR and FT-MIR spectroscopy." Unpublished report obtained from Bruker Optics, Inc., Manning Park, Billerica, MA.

Yang, Qingling, Atanasov, Plamen, and Wilkins, Ebtisam (1998) "Development of needle-type glucose snesor with high selelectivity." *Sensors and Actuators B,* **46**, 249-256.

Home Treatment

CHAPTER 8

**Tissue Modification with Feedback: The Smart Scalpel
E.L. Sebern, C.J.H. Brenan, I. W. Hunter**

Tissue modification with feedback: the "Smart Scalpel"

E. L. Sebern, C. J. H. Brenan, and I. W. Hunter

Department of Mechanical Engineering, Massachusetts Institute of Technology

Cambridge, MA 02139

ABSTRACT

While feedback control is widespread throughout many engineering fields, there are almost no examples of surgical instruments that utilize a real-time detection and intervention strategy. This concept of closed loop feedback can be applied to the development of autonomous or semi-autonomous minimally invasive robotic surgical systems for efficient excision or modification of unwanted tissue. Spatially localized regions of the tissue are first probed to distinguish pathological from healthy tissue based on differences in histochemical and morphological properties. Energy is directed to only the diseased tissue, minimizing collateral damage by leaving the adjacent healthy tissue intact. Continuous monitoring determines treatment effectiveness and, if needed, enables real-time treatment modifications to produce optimal therapeutic outcomes. The present embodiment of this general concept is a microsurgical instrument we call the Smart Scalpel, designed to cause hair growth delay or permanent hair removal.

1. BACKGROUND AND MOTIVATION

1.1 Feedback control

Feedback control (Figure 1) is widespread throughout many engineering fields, such as manufacturing, robotics, and in other human-machine interfaces. Feedback control uses measurement of the system output to modify the input in real-time. This on-line measurement strategy is necessary due to the difficulty of creating a comprehensive model to accurately predict the system output based on the input parameters. Feedback control provides a means to quickly respond to changes in the physical system or perturbations in the environment.
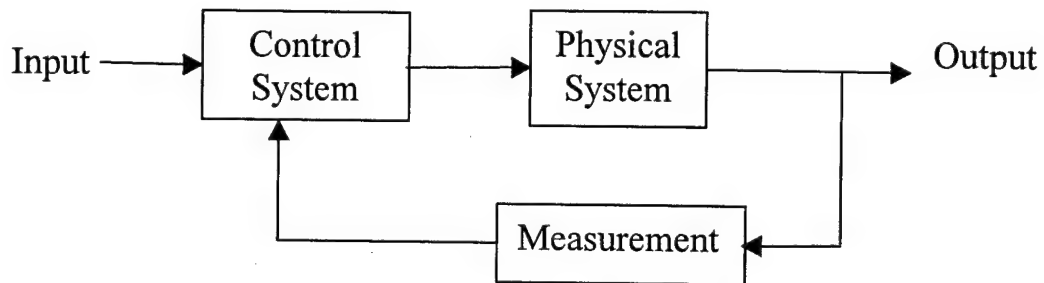
Figure 1: Illustration of classical feedback control loop in which the control system uses real-time measurement of the output to control the input to the physical system.

1.2 Application to medicine- the Smart Scalpel

An interesting application of feedback control is in the field of microsurgery. Many microsurgical procedures require a high degree of physical dexterity, accuracy, and control, which degrades rapidly with physician fatigue. This problem could be partially alleviated through inclusion of low-level decision-making embedded in a microsurgical instrument to aid in tissue location and removal. Our embodiment of this concept is an instrument we call the Smart Scalpel (Figure 2).
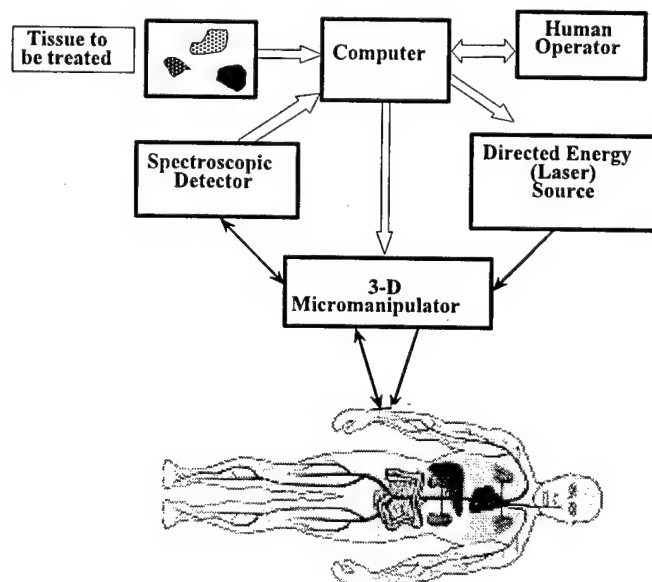

Figure 2: Schematic illustration of the "Smart Scalpel."

2

Implementation of the Smart Scalpel design is quite general in both measurement and intervention techniques. The human operator and computer model supply the computer with information regarding the properties of both normal and diseased/damaged tissue. The computer compares this information with real-time feedback of the histochemical and morphological properties of the tissue. Possible measurement techniques include: optical reflectance spectroscopy, magnetic resonance imaging, fluorescence optical spectroscopy, Raman spectroscopy, optical polarization detection, fluorescence polarization detection, mechanical impedance measurements, and electrical impedance measurements. The computer uses this feedback signal to identify the targets for a directed energy source, which affects only the diseased or damaged tissue and leaves the healthy tissue intact. The Smart Scalpel can employ a wide range of directed energy sources including: photon beam, electron or proton beam, localized electrical field, directed acoustic energy, and inertial cutting (low frequency mechanical energy). A micromanipulator serves as the interface between the patient and the imaging/therapeutic systems.

The many desirable attributes of the Smart Scalpel have the potential not only to improve the surgical performance in current microsurgical procedures but may yield the possibility of new surgical procedures not yet feasible with existing technology. The accuracy and reliability of present-day procedures may be enhanced and collateral damage minimized through an effective combination of analysis to discriminate between tissue types (e.g. diseased versus healthy) and removal/modification of the targeted tissue with a directed energy source. The Smart Scalpel diagnostics provide quantitative, rapid, on-line assessment of the procedure efficacy. This system of real-time feedback has great potential to increase patient comfort, shorten patient recovery times, and decrease the overall cost per procedure. Additionally, the Smart Scalpel is amenable to integration into a tele-operation system for remote surgery.

2. SMART SCALPEL APPLICATION TO HAIR REMOVAL

2.1 Current Clinical Practice

One application for the Smart Scalpel is in the area of permanent hair removal. Current methods for temporary hair removal include: shaving, cold or hot wax epilation, and chemical depilatories that often cause contact dermatitis.[1] Electrolysis is a permanent hair removal technique, but this method is tedious and only partially effective. Regrowth rates of 15% to 50 % have been associated with electrolysis.[2]

Because these hair removal techniques do not produce optimal therapeutic outcomes, laser hair removal has been explored. Two methods of laser hair removal have been tested with varying success. The first makes use selective photothermolysis to destroy or damage hair follicles using a normal-mode ruby laser (694 nm, 100-600 kJ/m^2, 270 μsec pulse width, 5-10 mm diameter spot). There are two dominant chromophores in skin that absorb this laser energy, melanin and oxygenated hemoglobin. Melanin is a chromophore present in the hair shaft or follicles, or both, which is absent in the dermis surrounding these follicles.[3] In the band from 650 nm to 700 nm, melanin absorption strongly dominates oxyhemoglobin absorption.[4] These longer wavelengths also penetrate deeply into the skin (550 μm to 750 μm) to selectively heat hair follicles in the underlying dermis.[5] Six month and two year follow-up studies of this normal-mode ruby laser treatment revealed that of the thirteen subjects tested, all laser exposures caused a hair growth delay, while four of the thirteen caused permanent hair removal of more that 50% of the treated region. The mechanism of laser hair removal is not well-understood, but histological studies from the normal-mode ruby laser study showed permanent hair removal correlated with miniaturization of the terminal hair follicles, rather than complete destruction of these structures.[6,7]

A second laser-based method of hair removal requires application of a carbon particle suspension, which fills the hair follicles, to selectively absorb energy from a Q-switched Nd:YAG laser. Mean percentage of hair regrowth at 1 month was 39.9%. At three months, the percentage of hair regrowth approximately doubled. The conclusion was that a single hair-removal treatment with the Q-switched Nd:YAG laser is safe and effective in delaying hair growth for up to 3 months.[8]

These laser treatments have drawbacks. In the normal-mode ruby laser treatment, great amounts of energy heat the skin regions without hair, making the procedure painful if the skin is not cooled. Currently a transparent material such as sapphire or glass is necessary to remove this heat from the epidermis.[3] Hair removal with both methods is not completely permanent. In the majority of cases, laser treatment only delays hair regrowth.

2.2 Smart Scalpel Approach

The more efficient Smart Scalpel approach is to first identify the hair follicles and target the laser to heat only these structures. This strategy leaves the tissue surrounding the follicles intact, so collateral damage is minimized. It is possible that by focusing laser energy on individual follicles, the Smart

Scalpel will be more effective in permanently removing hair, rather than delaying hair regrowth. Implementation of the Smart Scalpel strategy requires two elements: (1) a method to locate the hair follicles within the skin and (2) a means to direct the heating laser beam to the appropriate targets.

2.2.1 Spectroscopic identification of hair

Our approach to identify hair follicles makes use of selective absorption of light by the melanin present in the hair shaft and follicles. As mentioned earlier, melanin and hemoglobin are the two dominant chromophores of skin. A visible light reflectance spectrum of whole blood reveals high reflectivity of hemoglobin near 650 nm to 700 nm.[5] Melanin absorbs at this wavelength. Since melanin is present in the hair shaft and follicles, we can illuminate the follicle with 650 to 700 nm light and distinguish the absorbing hair follicles from the highly reflective blood vessels. Hemoglobin has a higher relative absorbance from 520 nm to 580 nm. Therefore, by taking the ratio of skin images illuminated with melanin-absorbing and hemoglobin-absorbing wavelengths, the hair signal can be clearly distinguished from the surrounding tissue. Melanin content in a dark human hair is 2.32% by weight, while human skin contains 0.023% and 0.008% by weight for dark and light-skinned patients, respectively.[9] Therefore, the strong melanin absorbance signal from hair can be distinguished from the melanin signal of other skin structures. Some image processing may be required to differentiate between the follicle and the rest of the hair.

2.2.2 Laser source

Once the follicles are identified, an appropriate laser source must be used to destroy these targets. We have identified several specifications for the laser source related to fluence, beam diameter interacting with the skin, laser size, and time-scale of laser-tissue interaction. Fluence levels depend on the wavelength of the laser used. In current treatments at 694 nm, the laser fluence is in the range of 100 to 600 kJ/m^2.[6,7] The wavelength of the laser light determines the coupling efficiency, which is the amount of energy applied to the tissue that is converted to thermal and/or mechanical energy. Therefore, if we use a laser wavelength, such as 1064 nm that melanin does not absorb as readily, we must increase the fluence of our laser.

We plan to illuminate the hair follicles with a 20 μm diameter laser beam, ~1/10 the diameter of a hair follicle. The common clinical practice in laser hair removal is to use a 5- 10 mm diameter beam.

Therefore, we can achieve the same fluence with laser energies ~10^5 times smaller than currently needed to destroy the follicles. A laser with a single spatial mode allows us to focus this beam to the small diameters required. Finally, our pulse width will determine whether the laser-tissue interaction is thermally-mediated or an adiabatic, mechanical process; this will be further developed in the following section. The size specification requires that the laser be small and lightweight so the physician and patient can comfortably interface with the SmartScalpel. If the laser cannot be made compact, we must be able to transmit the laser energy through a fiber optic so the laser is remote from the SmartScalpel.

2.2.2.1 Thermal Relaxation Time

In general, laser-tissue interactions can be grouped into two broad thermodynamic categories based on the time scale of the interaction. The dominant thermodynamic regime is characterized by the ratio of the laser illumination time, τ_{ill}, to the tissue thermal relaxation time, τ_r. For $\tau_{ill} \geq \tau_r$, the illuminated region is in thermal equilibrium with the surrounding tissue, and the tissue removal/transformation is primarily thermally mediated. When $\tau_{ill} < \tau_r$, laser energy is absorbed faster than it can be transported away from the illuminated region, and adiabatic processes determine the partitioning (and ultimate dissipation) of energy in the affected tissue volume. The thermal relaxation time for a cylindrical object, like a hair follicle, is given by:

$$\tau_r = \frac{d^2}{16\alpha} \qquad , \qquad (1)$$

$$\alpha = \frac{\beta}{\rho \cdot c} \qquad , \qquad (2)$$

where d is follicle diameter, and α is the thermal diffusivity, which is a combination of β, ρ, and c, the thermal conductivity, density, and specific heat, respectively.[10]

For tissue composed of 70% water, the material constants have the following values:

$$\beta = 4.21 \times 10^{-3} \frac{cm^2}{s}$$

$$\rho = 1.09 \frac{g}{cm^3}$$

$$c = 3.35 \frac{J}{g \cdot K}$$

Therefore, the thermal diffusivity, α, is 1.15×10^{-3} cm^2/s.[10] Assuming a 200 μm diameter blood vessel, which is typical for ectactic vessels in a port wine stain, the thermal relaxation time is approximately 20 msec.

Present laser-hair removal utilizes each of these two different thermodynamic regimes. The normal mode ruby laser, with longer pulse widths of 270 μsec relies on thermally-mediated processes to destroy the blood vessels. Alternatively, the Q-switched Nd:YAG laser, with pulse widths on the order of nanoseconds or picoseconds heats the carbon particles in a shorter timescale, destroying and/or damaging the hair follicle via an adiabatic, mechanical mechanism.

2.2.2.2 Light Absorption and Scattering

As laser light propagates through tissue, its intensity is attenuated by absorption and scattering. Beer's Law describes the amount of light attenuated by a tissue:

$$\frac{P_{out}}{P_{in}} = e^{-\mu_t x} \qquad (3)$$

where P_{in} is the power delivered to the tissue, and P_{out} is the power that passes through the tissue, in other words, the power that is not absorbed or scattered. The parameter μ_t is the optical extinction coefficient with the dimensions of [1/L], and x is the length over which the light interacts with the tissue. The mean free path is the value of x equal to $1/\mu_t$. Over the distance of one mean free path, the ratio of P_{out} to P_{in} is e^{-1}, so that P_{out} is approximately 35% of P_{in}. The optical extinction coefficient, μ_t, is given by:

$$\mu_t = \mu_a + \mu_s \quad , \quad\quad\quad\quad (4)$$

where μ_a and μ_s are the absorption coefficient and scattering coefficient, respectively.

The scattering coefficient determines how much of the light originating at the surface of the tissue actually reaches the structure of interest. When the characteristic dimension of scattering particles is much less than the wavelength of light passing through the tissue, Rayleigh scattering describes a relationship between scattering and wavelength. In this case, the power lost to scattering, P_s, is proportional to $1/\lambda^4$.

The absorption coefficient governs the amount of non-scattered energy that is absorbed by the structure,[11] which is a hair follicle in our application. The absorption coefficient varies with wavelength, and every material has its characteristic absorption spectrum. Melanin has much stronger relative absorption from 650 to 700 nm than oxyhemoglobin. Current normal-mode ruby laser therapy makes use of this selective heating for permanent hair removal. The absorption of melanin is much higher at shorter wavelengths, especially in the ultraviolet band, which would make shorter wavelength lasers more effective in heating the hair follicles. However, the amount of scattering at this wavelength is also substantially higher. The relative scattering for the 694 nm normal-mode ruby laser versus a 248-nm krypton-fluoride laser results in penetration depths of 750 μm and 2 μm, respectively.[5] Since hair follicles are approximately 500 μm below the skin surface, these longer wavelengths are required.

2.2.3 Microchip Laser

Given that the Smart Scalpel must comfortably interface with the physician and patient, the small microchip Nd:YAG laser, developed at MIT's Lincoln Laboratory may be a useful source for permanent hair removal. While the normal-mode ruby laser used in currently seems to yield the best clinical results, the strategy of focusing the laser to a smaller spot and spatially selecting the hair follicles in advance may require different laser parameters. Although melanin absorption is ~5 times less at the Nd:YAG wavelength, greater fluence can be delivered to these hair follicles because the rest of the skin is avoided. The 1064 nm wavelength also has greater penetration depths of approximately 1600 μm, so that deeper hair follicles may be destroyed/modified. Although the pulse width of the Q-switched laser

is much shorter than the thermal relaxation time of a hair follicle, a thermally-mediated process may not be required to cause permanent hair removal. The mechanism of laser hair removal is not understood, so an adiabatic, mechanical laser-tissue interaction may be just as effective or more effective in permanent hair removal.

The microchip laser is a passively Q-switched, single-mode, diode-pumped, Nd:YAG laser.[12] The Nd:YAG microchip laser array is fabricated with a thin, wide, resonator structure and is pumped by a two-dimensional diode laser array.[13] The diode-laser pump radiation is carried to the microchip via optical fiber. The primary advantage of using the microchip laser is that this source can be fit into a hand-held instrument that comfortably interfaces with the physician and patient. The laser beam has a single spatial mode (TEM$_{00}$) and was focused to a diffraction-limited, 5 μm diameter spot. We have tested two infrared lasers with the following specifications:

Microchip Laser	Wavelength (nm)	Pulse Energy (μJ/pulse)	Fluence (J/cm^2)	Pulse Width (psec)	Peak Power (kW)	Irradiance (GJ/cm^{2}s)
1	1064	120	610	450	267	1360
2	1064	210	1070	700	300	1530

In experiments with mouse skin, the laser converts a dark, absorptive hair into a more highly reflective and/or scattering material. One possible explanation is the hair pigment is photobleached by the high peak powers inherent to these short laser pulses. Another possibility is an increase in surface scatter through modification of the hair surface finish. An example of this effect with from a 1064 nm wavelength laser (120 μJ/pulse, 450 psec pulse width) is shown below. When the same microchip laser beam was focused on a hair follicle, the hair snapped away from the skin leaving little or no surface debris. These observations suggest that the microchip laser may be useful for hair removal.

Another interesting observation was that the 1064 nm beam caused a plasma to form at the laser focus. To better understand whether plasma formation is reasonable for this laser-tissue interaction, we estimate the electric field strength using the following derivation. In general, we calculate the electric field strength in terms of the irradiance of the laser and the tissue material constants. First, the Poynting vector, S, which is analogous to irradiance is given by the following equation:

$$S = \frac{1}{\mu_o} EB \qquad , \qquad (5)$$

where μ_o is the magnetic permeability in free space, ($4\pi \times 10^{-7}$ Ns^2/C^2), E is the electric field, and B is the magnetic field. The magnitude of S is the power per unit area crossing a surface whose normal is parallel to S.
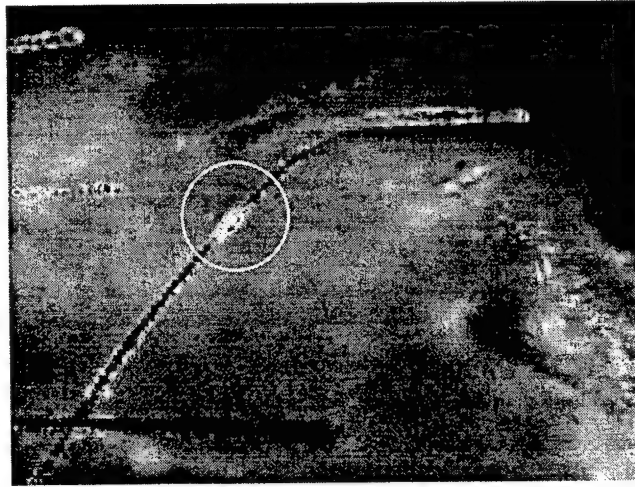

Figure 3: Image of mouse hair modified by 1064 nm microchip laser.

The E and B fields generated by a particle traveling through free space are perpendicular to each other and related by:

$$E = cB \qquad , \qquad (6)$$

where c is the speed of light (3×10^8 m/s). Using Maxwell's equations, the velocity is given by:

$$c^2 = \frac{1}{\varepsilon_o \mu_o} \qquad , \qquad (7)$$

where ε_o is the electric permittivity in free space, (8.8542×10^{-12} C^2/Nm^2), and, as mentioned above, μ_o is the magnetic permeability in free space. Equations 5, 6, and 7 can be manipulated to give the following expression for the Poynting vector:

$$S = \sqrt{\frac{\varepsilon_o}{\mu_o}} E^2 \qquad , \qquad (8)$$

For light propagating through a medium, one must account for the differences in permittivity and permeability in the medium versus these values in free space. These values are related through the absolute index of refraction, n:

$$n = \frac{c}{v} = \sqrt{\frac{\varepsilon \mu}{\varepsilon_o \mu_o}} \qquad , \qquad (9)$$

where ε is the permittivity of the medium, and μ is the magnetic permeability of the medium. For water, which composes 70% of tissue, at 20° C, the refractive index is 1.333. The highest peak power laser had an irradiance of 1.530×10^{16} W/m^2. Using these numbers, we calculate the field strength as,

$$E = \sqrt{\frac{S \mu_o c}{n}} \qquad , \qquad (10)$$

$$E = 2.08 \times 10^9 \frac{V}{m}$$

In the ablation of brain tissue, the laser tissue interaction with a peak irradiance of 5.7×10^{15} W/m^2 is reported to be plasma-mediated.[14] Therefore, it is highly probable that this magnitude of electric field strength leads to what we discerned to be plasma breakdown in our experiments.

3. SMART SCALPEL SYSTEM DESIGN

3.1 Prototype system

A diagram of the optical layout for our Smart Scalpel prototype is shown in Figure 4. The desired resolution for the imaging system is 20 μm, $\sim$ 1/10 the diameter of a hair follicle. Two LEDs provide red (660 nm, 10-16 W/sr) and green (565 nm, 0.44-0.63 W/sr) illumination. Each of the two LED outputs is collimated with a convex lens and made colinear with a dichroic beamsplitter. The two beams are then focused and relayed with a biconvex lens to one end of the fiber bundle. Between the lens and the fiber bundle, the light follows a path through a polarizing beamsplitter and a scanning galvanometer. The polarizing beamsplitter is used to pass the light of one polarization for illuminating the tissue and

reflect the orthogonal polarization of light backscattered from the skin to the photodetectors, which makes the underlying hair follicles more apparent. The imaging system galvanometer scans the light onto one end of the fiber bundle.
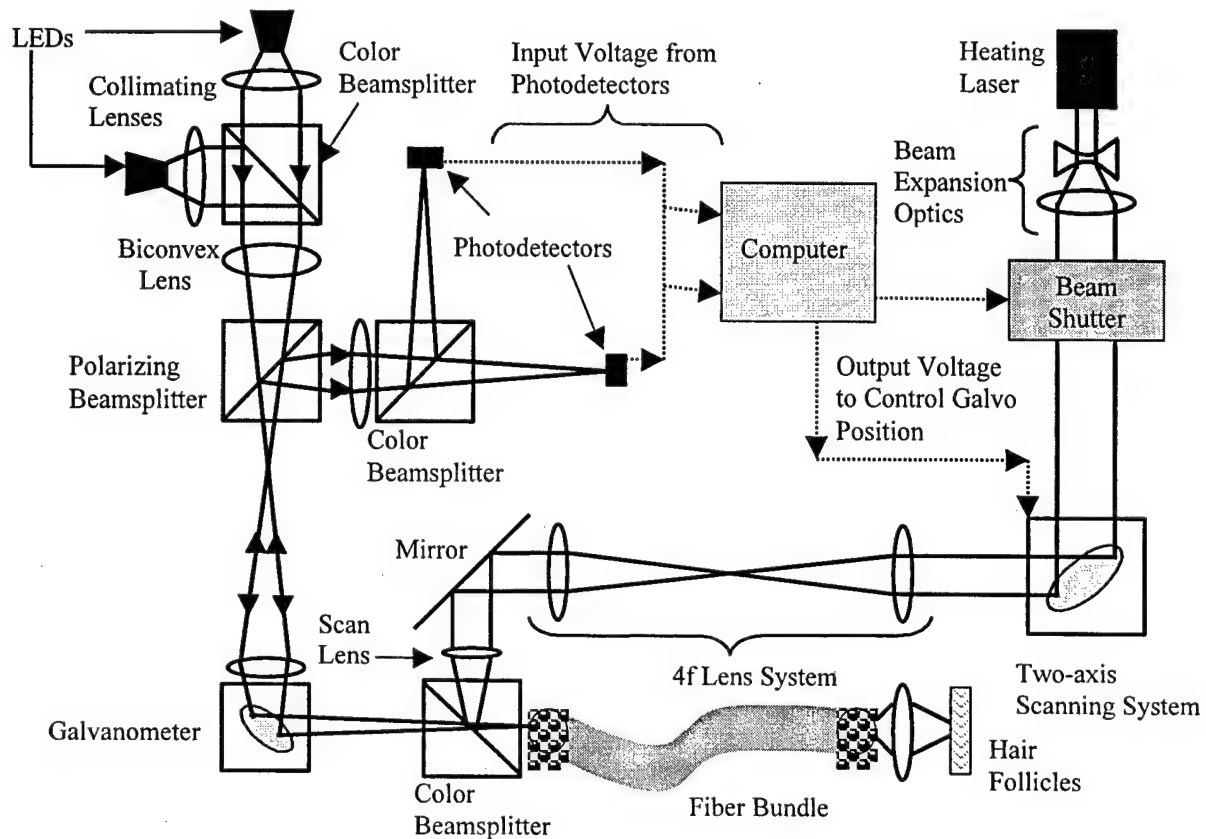
Figure 4: Schematic representation of Smart Scalpel beam scanning and imaging systems.

The light backscattered from the skin is transmitted through the fiber bundle, de-scanned by the galvanometer and reflected by the polarizing beamsplitter. A series of two biconvex lenses transmit the light reflected from the skin surface to two photodectors. A dichroic beamsplitter is used to separate the light into red (melanin-absorbing) and green (hemoglobin-absorbing) wavelengths. Each of these wavelengths is transmitted to a photodetector.

A data acquisition board converts the analog voltage outputs of these two arrays to digital signals, which are used by the computer to identify the spatial locations of hair follicles. The computer runs the necessary image processing algorithms on the array signals to distinguish hair follicles from the rest of the hair shaft and other melanin-rich structures. Once these coordinates are identified, the computer

controls the heating laser beam x-y position via a two-axis galvanometer scanning system. The minimum step response for the current galvanometers is ~ 1 ms, which is comparable to the thermal relaxation time, τ_r. Therefore, during the transit time between targets, a shutter blocks the laser beam to prevent heating of the tissue between hair follicles.
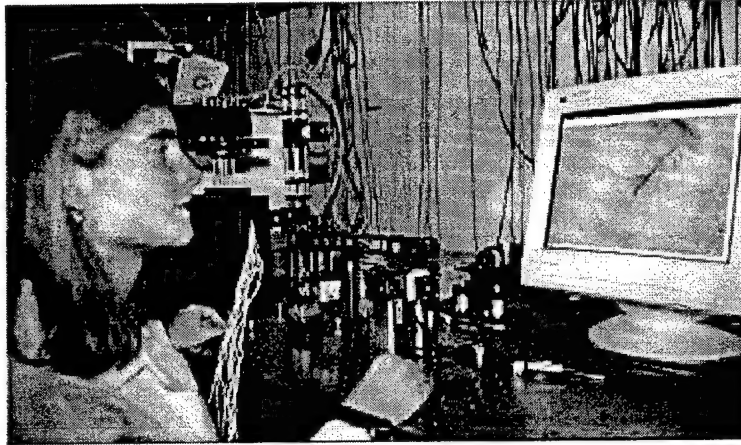

Figure 5: Photograph of prototype Smart Scalpel system.

The heating laser delivery subsystem of the Smart Scalpel is flexible in that any laser can be used with the system. The laser beam is first expanded to the maximum beam diameter that can be scanned by the mirrors on the two-axis galvanometer system. The two galvanometers provide random-access x-y spatial positioning of the laser beam. A 4F lens system is used to transmit the telecentric location between the two galvanometers to the final scan lens, which converts the beam rotation to a displacement scanned on the surface of the fiber bundle. This laser light is then transmitted to the skin surface through the fiber bundle. The final laser spot is approximately 20 μm to provide the desired spatial resolution for the 200 μm hair follicles.

3.2 Control Strategy

Many strategies may be used to deliver this laser energy to the hair follicles. One option is a point detection strategy in which a hand-held instrument is scanned across the skin surface. When a hair follicle is detected, the laser fires at the target. The key requirement for this strategy is that there be minimal time delay between the detection and energy delivery so as the physician scans the surface, the laser beam hits the correct targets. We assume that a physician scans the area at a rate of 10 mm/sec,

and we specify that the laser energy must be delivered within 20 μm ($D_{follicle}$/10) of the detected target. From these requirements, the follicles must be identified and treated within 2 ms.

A second strategy is a stationary device that rests on the skin surface and has a larger field of view than a single hair follicle. With this approach, targets can be identified in advance of the energy delivery. This may be necessary if image processing is required to distinguish the follicle from the rest of the hair. An important consideration for this full field approach is to minimize relative motion between the skin and the Smart Scalpel in the time to image, identify the target coordinates, and steer the laser beam to these locations. The maximum number of targets that can be addressed in one scan can be expressed as:

$$n = \frac{T_{move}}{(T_{exposure} + T_{acquisition} + T_{computer} + T_{galvos} + T_{ill})} , \qquad (5)$$

where n is the maximum number of targets per scan, and T_{move} is the period of the highest frequency component that could cause relative motion between the Smart Scalpel and the skin. Time delays in the Smart Scalpel feedback loop include: $T_{exposure}$, the integration time of the line array, $T_{acquisition}$, the time required to acquire and convert the photocurrents to a voltage output, $T_{computer}$, the data acquisition and processing time of the computer, T_{galvos}, the step response of the two-axis scanning system, and T_{ill}, the time required for the laser to thermally or mechanically treat the hair follicles. The highest frequency movement that has been identified is tremor. Tremor is classically said to be a 10 Hz quasi-sinusoidal displacement, although the frequency of tremor varies among different body parts and different people.[9] This movement requires that the region of skin be scanned and treated within ~100 ms.

3.3 Miniaturization

Once the prototype system is tested and the optimal control strategy is determined, the Smart Scalpel will be miniaturized to interface more comfortably with the physician and patient. If a point-detection strategy is used, our design can be implemented as a hand-held surgical instrument, which the physician scans over the skin surface (Figure 6). Two light-emitting diodes (LEDs) at desired wavelengths (i.e. 565 nm and 650 nm) and one or two photodetectors are mounted within the instrument. If LEDs are used to illuminate the skin, we could turn each LED on and off 180 degrees out of phase. With this method, a first reading from the photodetector would be acquired for a wavelength of high melanin absorption. A second photocurrent measurement would then be taken with the skin illuminated at the

565 nm wavelength to normalize the absorption measurement. We could utilize analog and/or digital circuitry to compare the photodetector currents and determine whether the instrument is located above a hair follicle or surrounding tissue. We must investigate the feasibility of this approach by determining the maximum time delay between initialization of the detection procedure and delivery of laser energy. If we want to direct the laser beam within 20 μm of the detected target, this delay cannot be longer than 2 msec, assuming a scan rate of 10 mm/sec.

Figure 6: Schematic of miniaturized Smart Scalpel.

As hair follicles are detected, laser energy is delivered to the instrument via a single mode fiber optic. If the microchip laser is used, the laser can be mounted within the instrument. A ball lens contacts the skin surface and focuses the laser light to the hair follicle. The ball lens also collects the reflected light, which is directed to the photodetectors.

4. FUTURE WORK

Project deliverables for the Smart Scalpel system focus on characterizing the spectroscopic sensitivity of the system and then testing a wide range of control strategies, lasers, and scan areas to determine the optimal treatment parameters.

- The spectroscopic sensitivity of the prototype system will first be evaluated with test patterns of hair. The optical reflectance measurements, the computer will identify the spatial coordinates of the hair and control the laser scanning system to cover only the identified targets.

- When the system can accurately recognize the hair and scan the laser over only these coordinates, the next step will be to test the imaging and treatment components of the system using an animal model. Through these experiments, much will be learned about the laser wavelength, fluence, pulse width, control strategy, and other parameters leading to the optimal therapeutic outcome.

- The final deliverable for the hair removal project is to develop a robust system that interfaces well with both physician and patient in a clinical setting.

REFERENCES

1. R.N. Richards, U. Marguerite, and G. Meharg, "Temporary Hair Removal in Patients with Hirsutism: A Clinical Study," *Cutis*, 45, pp. 199-202, 1990.

2. R.F. Wagner, "Physical Methods for the Management of Hirsutism," *Cutis*, 45, pp. 319-26, 1990.

3. R.R Anderson, "Laser Medicine in Dermatology," *Journal of Dermatology*, 23(11), pp. 778-782, November, 1996.

4. J.B. Dawson, D.J. Barker, D.J. Ellis, E. Grassam, J.A. Cotterill, G.W. Fisher, and J.W. Feather, "A Theoretical and Experimental Study of Light Absorption and Scattering by *in vivo* Skin," *Phys. Med. Biol.*, 25(4) pp. 695-709, 1980.

5. R.R. Anderson, and J.A. Parrish, "The Optics of Human Skin," *The Journal of Investigative Dermatology*, 77, pp. 13-9, 1981.

6. M.C. Grossman, C. Dierickx, W. Farinelli, T. Flotte, and R.R. Anderson, "Damage to Hair Follicles by Normal-Mode Ruby Laser Pulses," *Journal of the American Academy of Dermatology*, 35, pp.889-894, 1996.

7. C. Dierickx, M.C. Grossman, W. Farinelli, and R.R. Anderson, "Permanent Hair Removal by Normal-Mode Ruby Laser," *Arch. Dermatol.*, 134, pp. 837-842, 1998.

8. C.A. Nanni, T.S. Alster. "Optimizing treatment parameters for hair removal using a topical carbon-based solution and 1064-nm Q-switched neodymium:YAG laser energy," *Arch. Dermatol.*, 133(12), pp. 1546-9, December 1997.

9. K.P. Watts, R.G. Fairchild, D.N. Slatkin, D. Greenberg, S. Packer, H.L. Atkins, and S.J. Hannon. "Melanin Content of Hamster Tissues, Human Tissues, and Various Melanomas," *Cancer Research*, 41(2), pp. 467-472, February 1981.

10. A.L. McKenzie, "Physics of Thermal Processes in Laser-Tissue Interactions." *Phys. Med. Biol.*, 35(9), pp. 1175-1209, 1990.

11. W.F. Cheong, S.A. Prahl, and A.J. Welch, "A Review of the Optical Properties of Biological Tissues," *IEEE Journal of Quantum Electronics*, 26(12), pp. 2166-2185, 1990.

12. R.S. Afzal, A.W. Yu, J.J. Zayhowski, and T.Y. Fan, "Single Mode, High Peak Power, Passively Q-Switched Diode-Pumped Nd:YAG Laser," *Optics Letters*, 22(17), pp. 1314-1316, 1997.

13. C.D. Nabors, J.J. Zayhowski, R.L. Aggarwal, J.R. Ochoa, J.L. Daneu, and A. Mooradian, "High-Power Nd:YAG Microchip Laser Arrays." *Optical Society of America Proceedings on Advanced Solid-State Lasers,* 13, Proceedings of the Topical Meeting, pp. xvii+391, 234-6, 1992.

14. J.P. Fischer, J. Dams, M.H. Gotz, E. Kerker, F.H. Loesel, C.J. Messer, M.H. Niemz, N. Suhm, J.F. Bille. "Plasma-Mediated Ablation of Brain Tissue with Picosecond Laser Pulses." *Applied Physics B.* 58, pp. 493-499, 1994.

Home Treatment

CHAPTER 9

**Progress in the Development and Application of Dynamic Compliance Spectroscopy
for Early Detection and Prevention of Pressure Ulcers**

C.J.H. Brenan, B. Sebern, I. W. Hunter

Progress in the development and application of dynamic compliance spectroscopy for early detection and prevention of pressure ulcers.

Colin J.H. Brenan, B. Sebern and I.W. Hunter
Department of Mechanical Engineering
Home Automation Consortium
Massachusetts Institute of Technology

1.0 Introduction

A pressure ulcer is any lesion caused by unrelieved pressure resulting in damage of underlying tissue (Abruzzese, 1985). Decreased blood flow to tissue in prolonged contact with a support structure (bed or chair) or body coverings (e.g. clothes, bed sheets or blankets) precipitates formation of a pressure ulcer in the contact region. Consequently, at greatest risk are patients immobilized for an extended period of time. Early diagnosis and intervention is therefore important to prevention of pressure ulcers and containment of health costs associated with their treatment.

An increasingly large sub-group of the at-risk patient population are elderly. There are several risk factors that make elderly patients particularly vulnerable to pressure ulcers. As an individual ages, the increased susceptibility to debilitating conditions (disease, broken bones, etc.) can severely curtail mobility and force the individual to spend long periods of time in a bed or chair. When combined with the physiological degradation of tissue associated with aging (e.g. blood circulation, mechanical compliance and resiliency), the likelihood of a pressure ulcer occurring increases dramatically. Pressure ulcers can be the primary disease condition requiring treatment or, as is more often the case, a complication to the primary disease condition for which the individual is receiving treatment.

A cause for concern is the high incidence and prevalence of pressure ulcers in hospitalized and nursing home populations as well as among persons receiving home healthcare. The incidence of pressure ulcers in hospitals has been found to be as high as 30% (Gerson, L.W., 1975) while the percentage of individuals diagnosed with pressure ulcers in skilled care and nursing home-type facilities was somewhat less (23%) (Brandeis *et al.*, 1990). At substantially greater risk are sub-populations of hospitalized persons

whose physical affliction limits movement. Sixty percent of quadriplegic patients are found to have pressure ulcers while elderly patients immobilized during the healing of a femoral fracture had a sixty-six percent occurrence of pressure ulcers. Thirty-three percent of critical care patients were found to suffer from pressure ulcers. The prevalence among persons cared for in home settings with supervision or assistance of professionals is not fully understood because there is little research on the subject (Barbenel *et al.*, 1977).

Commensurate with its enormous health impact, treatment costs of pressure ulcers are high and are anticipated to increase in the future as the population ages and as a larger percentage is hospitalized (Miller and Delozier, 1994). For the 34,000 cases of pressure ulcers diagnosed in hospitals for 1995, the average cost of 31.5 k$ per case resulted in an annual cost of 836 M$. Similarly, as a complication stemming from long-term immobilization during the healing of hip fractures, treatment of pressure ulcers extracted a cost exceeding 84 M$ from the health care system. Exceeding 1335 M$ annually, the cost to treat elderly patients with pressure ulcers either at home or in nursing homes was by far the greatest. The annual cost for treatment of pressure ulcers is greater than 2100 M$ or approximately 1% of total healthcare expenditures.

Pressure ulcers usually occur over bony prominences and are graded or staged to classify the degree of tissue damage observed. The present diagnostic methodology relies primarily on visual inspection and gentle palpitation of the afflicted area to assess the degree of tissue degradation [AHCPR, 1996; Bergstrom *et al.*, 1987). The classification scheme for pressure ulcer assessment begins with Stage 0 corresponding to normal healthy skin having no redness or break over a bony prominence. Stage 1 is characterized by a nonblanchable erythema (abnormal redness of skin due to capillary congestion) is observed in the intact skin over a bony prominence (Figure 1). In Stage 2, there is a partial loss in skin thickness involving the epidermis and/or dermis layers. Tissue damage in these early stages (Stage 1 or Stage 2) is reversible if the pressure ulcer is detected and an appropriate therapy applied. This is not true for later stage pressure ulcers.

A Stage 3 pressure ulcer is when the epidermis and dermis skin layers are broken, exposing the underlying subcutaneous tissue. The wound edges are distinct, perhaps

2

undermined and there may also be necrotic tissue present. Stage 4, the most severe and advanced form of a pressure ulcer, is characterized as a break in the skin (wound) extending through the underlying tissue and subcutaneous layers to expose the muscle or bone. There is substantial skin loss, tissue necrosis and damage to muscle, bone or supporting structures (e.g. joint capsule or tendon). Treatment of these deep wounds becomes complicated requiring hospitalization and intensive nursing care.
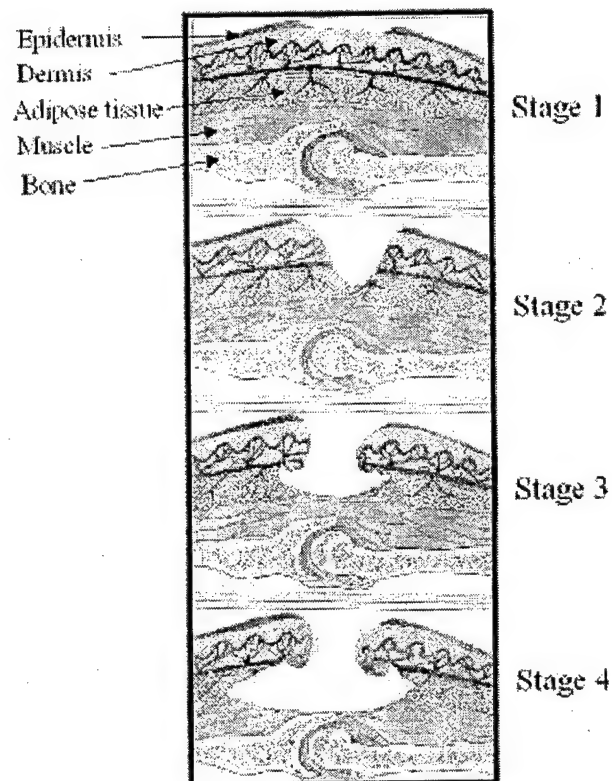
Figure 1: Classification scale for pressure ulcer assessment (Sebern, 1987).

There are several problems to the present assessment procedure. Visual observation of skin surfaces characteristic of Stage 1 and 2 pressure ulcers exhibit few distinct cues to ascertain subsurface mechanical degradation of tissue. Indeed, identification of Stage 1 pressure ulcers becomes particularly acute with dark-skinned patients. Second, the classification scheme is qualitative thereby making quantitative assessment and comparison of pressure ulcers under variable skin conditions difficult. For example, accurate staging is

3

not possible when a scab is present. Furthermore, casts and other orthopedic paraphernalia can partially occlude the ulcer thus increasing assessment difficulty.

2.0 Research Plan and Approach

2.1 Sensor modalities

An objective of this study is to discover a sensor modality or combination of sensor modalities whose outputs are combined in such a way as to unequivocally identify and isolate the region of tissue degradation defining a pressure ulcer. Any viable modality must be either non- or minimally-invasive and capable of identifying the ulcerous region at an early stage (Stage 1 or Stage 2) with a high degree of certainty. Since the tissue degradation in these early stages is in the epidermal layer or below, the detection technique must be able to penetrate and analyze tissue properties below the skin's surface. Finally, the detection method must also be inexpensive and simple to learn and operate given it will be used in a home healthcare environment.

The first element of our research plan is to identify transduction methods potentially suitable for early, non-invasive detection of pressure ulcers. Degradation of the epidermal and dermal cellular matrices characteristic of early stage pressure ulcers result in changes in different tissue properties. Skin is a laminate, non-linear visco-elastic material; hence, microscopic changes in cellular mechanical properties, inter-cellular adhesion or adhesion between tissue layers are observable as macroscopic changes in one or more static or dynamic mechanical tissue properties. Thus, a probe of tissue mechanical properties could potentially tissue degradation resulting from early onset of a pressure ulcer.

Skin is also an electrical conductor; therefore, spatially resolved electrical impedance measurement (either static or dynamic) can potentially detect the position of a pressure ulcer below the skin. Pressure ulcers are also characterized by histochemical changes stemming from restricted blood flow to the afflicted area. Full-field spectral imaging could be an attractive method to visualize pressure ulcers based on their different spectrochemistry compared to the surrounding healthy tissue. One viable possibility is multi-wavelength reflectance spectral imaging for quantitative measurement of tissue

oxygenation levels. It is important to note that the output of two or more sensor modalities could be combined in such a manner as to enhance the probability of detecting a pressure ulcer.

2.1.1 Mechanical compliance spectroscopy

Review of the present classification scheme suggests to us a reasonable starting point for our investigations is to consider the skin's mechanical compliance as a function of tissue degradation correlate with Stage 1 & 2 pressure ulcers. An automatic mechanical probe of skin compliance is quite attractive because it compliments the visual observations of a trained observer as put forth in the standard ulcer classification scheme. This combination could be quite effective in the location and quantitative assessment of pressure ulcers. At a later stage of development, an automated machine vision could potentially replace the trained observer to fully automate the detection and analysis process.

An additional benefit of a mechanical measurement approach is the potential for not only ulcer diagnosis but also therapeutic intervention with the same system. Integration of mechanical compliance sensors in a bed enables continual monitoring of the mechanical compliance of skin in contact with the bed's surface. If the compliance measurement indicates the early beginnings of a pressure ulcer, the same system could be employed to modify the local tissue loading or apply a mechanical stimulus to enhance blood flow into the affected tissue.

2.2 Technological and mathematical tools

The second component of our research plan is the adaptation of the different tools (both mathematical and technological) we have developed for tissue modeling. We have built instrumentation for simultaneous, full-field stress-strain analysis of tissue deformed over a physiologically-relevant range of strain (up to 20%) (Charette and Hunter, 1997; Charette et al., 1997). We have extensive experience and knowledge in the application and interpretation of linear and non-linear (Korenberg and Hunter, 1996; Korenberg and Hunter, 1990; Hunter and Korenberg, 1986) system identification techniques applied to biological systems. The essence of system identification is to perturb the input to a system

in a known manner and observe the system output correlate with the input. In this manner a causal link is established between system input and output thus enabling physically plausible models describing system function to be constructed and evaluated in a rational manner. A typical system identification approach is to apply a time-varying input (e.g. sinusoidal or stochastic) to the system and measure the characteristic system response in time (impulse response) or frequency (transfer function).

2.3 Pressure ulcer detection system

After we have selected an effective sensor modality (or combination of modalities), the third component of the research plan is to integrate the sensor into a prototype system for pressure ulcer detection. We will identify the key design parameters important to development of a detection system usable outside a laboratory environment. Emphasis will be placed on creating a portable functional unit that is easy to handle and use. This process will occur in consultation with our corporate sponsor in the Home Automation consortium (Hill-Rohm) and other potential users of the technology (e.g. nurses, nurse practitioners). The prototype will be evaluated, first, with pressure ulcer tissue phantoms after which its operation will be assessed with ulcerous tissue.

3.0 Measurement Methodology

3.1 Mechanical Compliance Spectroscopy

Skin is a non-linear, viscoelastic material whose localized displacement to an applied force is a complex function of tissue microstructure and adhesion between the tissue layers comprising the skin. For small force perturbations about an applied static force, the skin mechanical dynamics can be reasonably approximated as a linear mechanical system relating the applied force $F(t)$ to skin deformation $x(t)$ as

$$F(t) = I\frac{d^2 x(t)}{dt^2} + B\frac{dx(t)}{dt} + K x(t) \qquad\qquad 1$$

where I, B and K are the lumped mechanical parameters describing the skin's mechanical inertia, viscosity and stiffness, respectively. Taking the Laplace transform of Equation 1

gives an equivalent transfer function representation of the skin's mechanical compliance as a function of frequency, ω

$$\frac{x(\omega)}{F(\omega)} = \frac{G\omega_n^2}{\omega^2 + 2j\zeta\omega_n\omega + \omega_n^2}$$

2

in which

$$G = \frac{1}{K}, \qquad \omega_n = \sqrt{\frac{K}{I}} \quad \text{and} \quad \zeta = \frac{1}{2}\frac{B}{\sqrt{IK}} \quad .$$

3

Measurement of the compliance transfer function for different static force loadings of the skin will enable us to estimate the skin inertia, viscosity and stiffness as a function of frequency and static force and correlate these results with the physiological state of the tissue. *A priori* one could expect a large change in these parameters with early stage pressure ulcers given the tissue delamination characteristic of these ulcers. Indeed, our previous work in tissue (muscle) physiology will also serve as a guide to interpretation of the results and in the design of the measurement approach. We have measured, for example, an increase by over a factor of ten in elastic stiffness (K) and an increase of a factor of five in viscous stiffness (B) between relaxed and stimulated whole skeletal muscle (Kearney and Hunter, 1990).

The mechanical properties of ulcerous skin are anticipated to be quite different from healthy skin. Tissue inertia (I) is a function of tissue mass in the probed region; thus, in reaction to the tissue damage from a pressure ulcer, tissue inertia may increase as the volume of fluid (water or blood) in the damaged tissue increases. Similarly, viscous stiffness (B) and elastic stiffness (K) could also be affected because of the non-linear mechanical behavior of porous fluid filled structures. Collagen detachment between adjacent skin layers (epidermis and dermis) would be measurable as a decrease in skin elastic stiffness.

3.2 Preliminary Results

We have begun assessing the application of mechanical compliance spectroscopy for non-invasive detection of pressure ulcers. To that end, we built an apparatus for measurement the mechanical impedance transfer function of skin tissue and tissue phantoms. Our input is a force perturbation applied to the skin's surface and our output is the resulting skin displacement. The ratio of displacement to applied force is the mechanical compliance. Measurement of the compliance as a function of the force perturbation frequency yields the compliance transfer function (Equation 2).
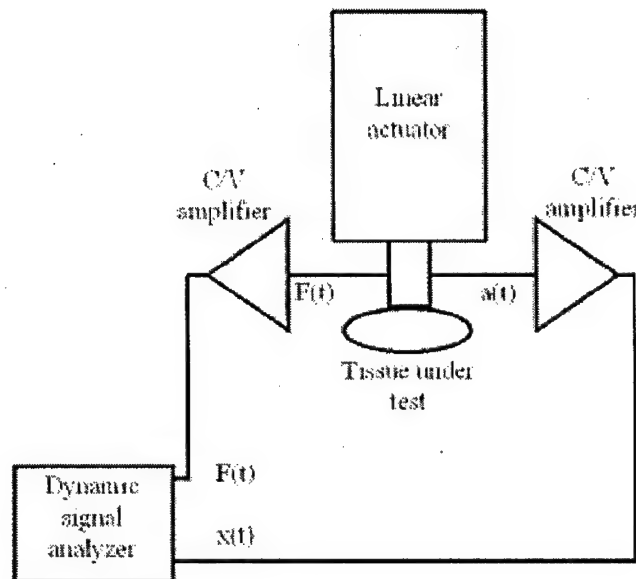
Figure 2: Diagram of apparatus for mechanical compliance transfer function measurement.

The prototype testing apparatus we built to begin our measurements is diagrammed in Figure 2 and a photo of the apparatus is in Figure 3. A linear electromagnetic actuator (Bruel and Kjaer 4910) is mounted vertically in a rigid frame and a piezoelectric accelerometer (Type 8001) is attached firmly to the actuator platform. The 4910 actuator generates a maximal force of 10 N with a maximal displacement of 6 mm over a wide range of frequencies (DC - 1000 Hz). The two outputs from the accelerometer simultaneously record the force applied to the tissue and its resulting acceleration. One

charge-to-voltage amplifier transforms the force output into a proportional voltage while a second amplifier doubly integrates the acceleration signal to give tissue displacement. The actuator is driven by the output from a dynamic signal analyzer (DSA) (HP 3562A) through a power amplifier. We used a swept sine approach in which the DSA outputs a time-varying sinusoidal voltage to drive the actuator at a different frequency within a range of pre-selected frequencies. The measured force and displacement signals are input to separate DSA channels which, in turn, calculates the compliance at each frequency generated by the DSA. The stimulus applied to the system can be stochastic, frequency swept sinusoids or some other function tailored to optimally analyze the mechanical dynamics of a particular system. A DC offset is added to the drive signal to apply a static force to the skin and the force perturbation amplitude was purposely kept substantially smaller ($< 1/10$) than the static force load.
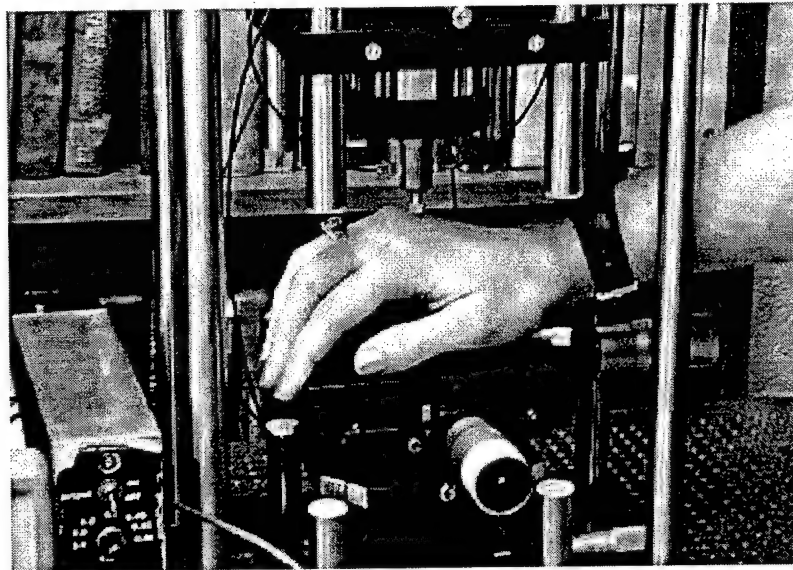
Figure 3: Photo of apparatus for mechanical compliance transfer function measurement.
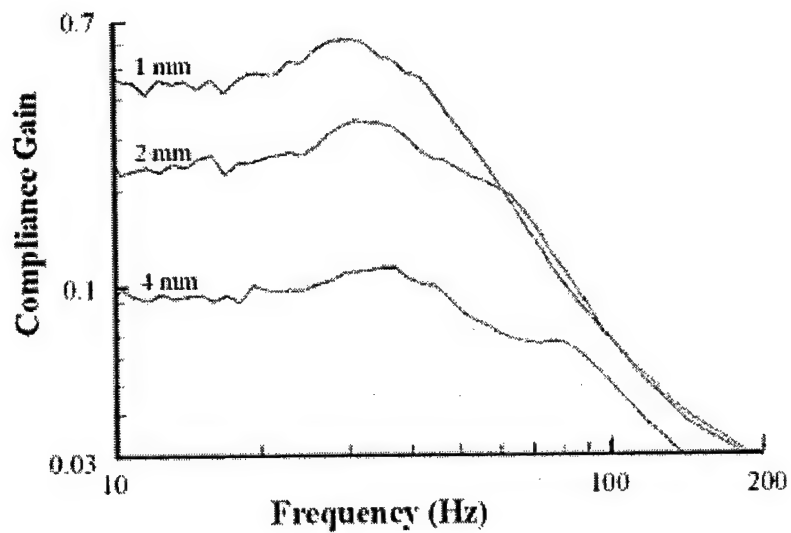
Figure 4: Gain portion of the compliance transfer function of forearm upper surface under three different static loads.
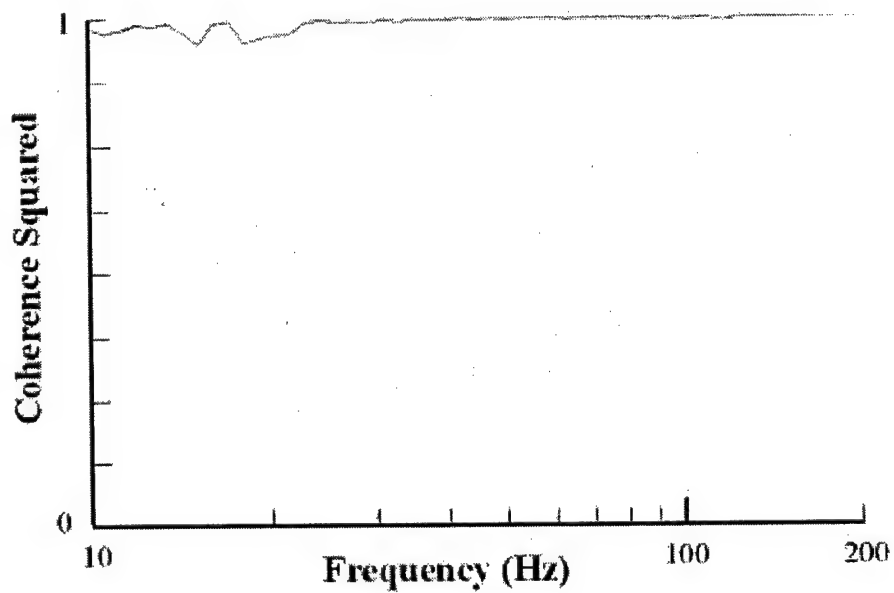

Figure 5: Coherence function for one of the compliance transfer function measurements.
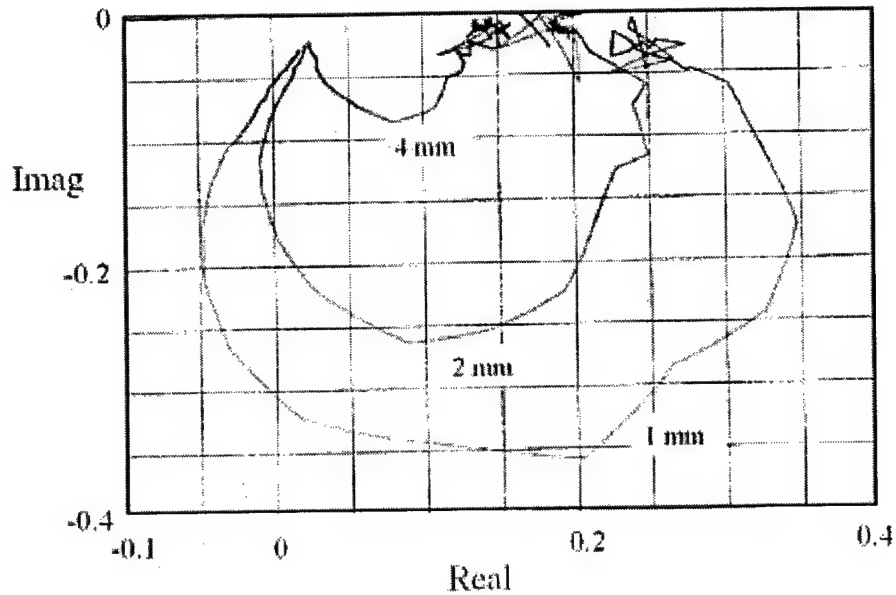
Figure 6: Nyquist plot showing real and imaginary components of compliance transfer function for three different static loads.

Preliminary results with this apparatus in the measurement of the compliance transfer function of normal human skin are shown in Figures 4 through 6. For this measurement set, the accelerometer was brought initially into light contact with the forearm on its upper surface approximately 100 *mm* from the wrist. Starting with no deformation of the skin (zero bias force), a static force was applied to the skin to deform it by 1, 2 and 4 *mm*. At each bias position the compliance transfer function was measured on application of a small force perturbation swept in frequency over a range from 10 to 200 *Hz*. Interestingly, at each static loading, the compliance transfer function behaves as a second-order mechanical system (Equations 1 & 2) with ω_n (~ 37 *Hz*) and a damping ratio ζ (~ 0.4) essentially independent of bias loading force (Figure 4). The transfer function gain ($G\omega_n^2$) scaled inversely with respect to static loading. A preliminary analysis indicates the skin's mechanical compliance is dominated by its stiffness K and to a much lesser extent inertial and viscous factors (I and B, respectively). The coherence function (Figure 5) for all transfer function measurements exceeded 0.95 across the measurement frequency range indicating the system is linear and therefore further validating our original observation and

11

assumption. Finally, since the transfer function is a complex quantity a Nyquist representation plotting the real and imaginary components of the measured transfer functions shows a clear separation between the three different static loading cases considered (Figure 6).

4.0 Future Plans

1. Develop a suitable tissue phantom to mimic Stage 1 & 2 pressure ulcers. Use as a test bed for exploring stimulation and measurement strategies for implementation of mechanical compliance spectroscopy.

2. Augment linear and non-linear system identification tools developed previously for analyzing tissue mechanical dynamics for application to interpretation and analysis of mechanical compliance spectra from ulcerous and healthy skin.

3. Relate mechanical parameters and observed mechanical dynamics to tissue microstructure and histochemistry to increase accuracy and efficacy of predictability from single point, lumped parameter measurement.

4. Explore and develop new diagnostic techniques to potentially combine with mechanical impedance spectroscopy for increased detection accuracy. Such techniques to consider include Doppler blood flow measurement (acoustic or optical), electrical impedance spectroscopy and tissue oxygenation from multi-spectral optical analysis.

5. Build and test a prototype hand-held unit suitable for pressure ulcer detection.

6. Explore other mechanical measurement concepts suitable for a wearable device or for integration into a bed. Some ideas include sensor arrays for distributed mechanical compliance testing and a wearable device for continuous monitoring and mechanical stimulation of pressurized tissue (similar to a Band Aid).

5.0 References

Agency for Health Care Policy and Research (AHCPR) Pressure Guidelines, National Institutes of Health; information found at their website http://text.nlm.nih.gov.

Abruzzese, R.S. (1985). Early assessment and prevention of pressure sores. In: Lee, B.Y., editor. Chronic ulcers of the skin. New York, McGraw-Hill; 1-19.

Barbenel, J.C., Jordan, M.M., Nicol, S.M. and Clark, M.O. (1977). Incidence of pressure sores in Greater Glasgow Health Board area. Lancet, 2 (8037), 548-550.

Bergstrom N., Braden B.J., Laguzza A. and Holman, V. (1987). The Braden scale for predicting pressure sore risk. Nurs. Res., 36 (4), 205-210.

Brandeis, G.H., Morris, J.N. Nash, D.J. and Lipsitz, L.A. (1990). The epidemiology and natural history of pressure ulcers in elderly nursing home residents. JAMA, 264 (22), 2905-2909.

Charette, P. and Hunter, I.W. (1997). Large deformation mechanical testing of biological membranes using speckle interferometry in transmission. I: Experimental apparatus. Applied Optics, 36, 10.

Charette, P., Hunter, P. and Hunter, I.W. (1997). Large deformation mechanical testing of biological membranes using speckle interferometry in transmission. II: Finite element modeling. Applied Optics, 36, 10.

Gerson, L.W. (1975). The incidence of pressure sores in active treatment hospitals. Int. J. Nurs. Stud., 12 (4), 201-204.

Hunter, I.W. and Korenberg, M. (1986). The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. Biol. Cybern., 55, 135-144.

Kearney, R.E. and Hunter, I.W. (1990). System identification of human joint dynamics. CRC Critical Reviews of Biomedical Engineering, 18, 55-87.

Korenberg, M. and Hunter, I.W. (1996). The identification of nonlinear biological systems: Volterra kernel approaches. Ann. Biomed. Eng., 24, 250-268.

Korenberg, M. and Hunter, I.W. (1990). The identification of nonlinear biological systems: Wiener kernel approaches. Ann. Biomed. Eng., 18, 629-654.

Miller H. and Delozier, J. (1994). Cost implications of the pressure ulcer treatment guideline. Columbia (MD): Center for Health Policy Studies. Contract No. 282-91-0070. Sponsored by the Agency for Health Care Policy and Research.

Sebern, M. (1987). Home-team strategies for treating pressure sores. Nursing87, April.

Sensors and Actuators

CHAPTER 10

Conducting Polymer Sensors for the Home
          P. Madden, J. Madden, T. Kanigan, I. W. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Conducting Polymer Sensors for the Home

Peter G. Madden, John D. Madden

Dr. Tanya Kanigan and Professor Ian W. Hunter

Introduction

Conducting polymers exhibit a wide range of tunable properties that make them ideal for use as sensors. Examples of applications are chemical, mechanical, optical, thermal, acoustic, and electrical sensors. Most of these sensor modalities involve changes in the electronic structure of the polymer resulting from chemical, mechanical, optical, thermal and acoustic stimuli. We have been focussing efforts on perfecting individual device components, namely transistors, force transducers, and displacement transducers, with the aim of integrating the functional elements into complete devices. Completed milestones include the demonstration of the first known polymer-based strain gage/ force transducer. Key characteristics include tunable gage factor (0.2 to 5), large recoverable strain (>1%) and low cost. Conventional strain gages, by comparison, have gage factors of 2 and recoverable strains of only 0.1%. Polyaniline transistors have been demonstrated and calculations are presented that indicate their suitability for use in strain gage amplifiers, pointing the way towards a low cost, integrated, all polymer device. Finally, some fundamental measurement techniques are described that enable the electrochemical characterization of the polymer devices. These techniques are fundamental to determining ultimate device efficiency and performance.

Strain gage results

In this section, we report on the experimental determination of gage factor in conducting polymer-based strain gages. These are, to our knowledge, the first measurements of

1

conducting polymer gage factor. Polymer strain gages are of great interest due to their large recoverable strains, low cost, and potential for integration with other polymer devices including transistors and batteries. This project will continue into the next phase, leading to the fabrication of an integrated strain gage, amplifier and power supply. We begin by providing an overview of strain gages and their use in measuring displacement and force, followed by a presentation of results. A discussion of polymer transistors further on in this text focuses on the use of these transistors to amplify the strain gage output.

Strain gages and force transducers: Background

Strain gages are widely employed to record displacement and force. They allow the measurement of small strains (typically 0.0001% to 0.1%) characteristic of the deformation of inorganic solids under tension. They may also be employed to form high stiffness load cells in which the recorded strains are related to stresses via elastic modulus. The discovery of conductive polymers creates the possibility of generating inexpensive, flexible, and tough gages, featuring recoverable strains that exceed those of metal and silicon-based gages by an order of magnitude. Furthermore, the prospect of co-fabricating strain gages along with batteries, and amplifiers, is attractive from the standpoints of cost and miniaturization.
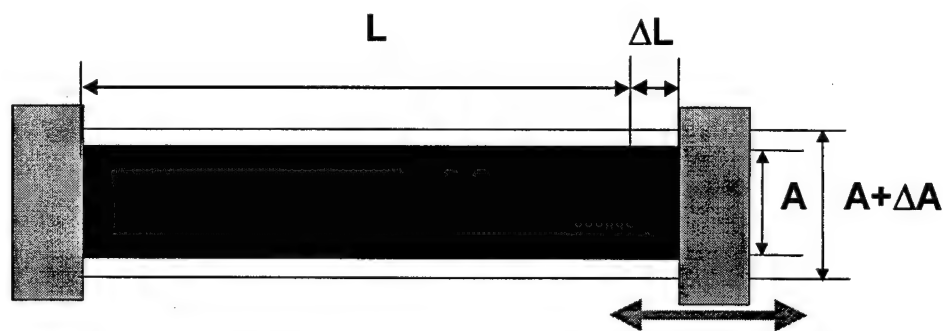
Strain gages rely on the change of resistance, R, induced in a conducting material as it is deformed. This change is recorded, and related back to strain, or, via the elastic modulus and the material geometry, to force. The ratio of relative change in resistance to strain,

known as the gage factor, K, is thus a key figure of merit for any material being considered for use in strain gages. The larger the gage factor, the more sensitive the strain gage. Two factors affect the relative change in resistance of a material as it is deformed, namely a geometry factor related to the effect of a change in material dimensions on resistance, and a piezoresistive factor, which results from changes in resistivity, ρ, with deformation (Figure 1). These factors and their relationships to gage

---

Figure 1: Gage Factor equations and diagram

$$K = \frac{\Delta R}{R\varepsilon} \qquad R = \frac{\rho L}{A}$$

$$K = \frac{\Delta \rho}{\rho \varepsilon} + 1 + \frac{\Delta A}{A\varepsilon} = \frac{\Delta \rho}{\rho \varepsilon} + (1 + 2\upsilon)$$
(Isotropic)

---

factor are summarized by the equations in Figure 1. In this equation, A represents the cross-sectional area normal to the strain. The right hand most expression applies only to isotropic materials. In isotropic materials the geometric contribution to gage factor is related to Poisson's ratio, ν, and hence ranges between 1.6 and 2. K values for metallic

strain gages are typically near 2.0, whereas silicon gages may range between 40 and 200, depending on doping levels. Maximum recoverable strains are approximately 0.1 %.

Two other factors affecting strain gage performance are the temperature coefficient of resistance, which is the ratio of relative change in resistance to temperature, and the material resistivity. While silicon-based strain gages are very sensitive to deformation, they are also extremely sensitive to temperature fluctuations, with relative changes varying between 300 and 4000 K^{-1}. Constantan gages exhibit temperature coefficients of 20. It is critical to monitor and or control the temperature of these strain gages, especially for long term measurements.

Resistivity is important because it determines the appropriate gage dimensions. The larger the resistance, the lower the current required, and hence the lower the power dissipation and Joule heating. Furthermore, one usually seeks to maximize the potential drop, in order to maximize the magnitude of the potential change. The low resistivity of metals (Cu 1.6 $n\Omega \cdot m$, Constantan 1.7 $\mu\Omega \cdot m$) means that very thin films or very long, wound, wires must be employed in order to generate a sufficient potential drop. These require packaging for mechanical stability, which reduces thermal contact with the environment and adds complexity and cost to the fabrication process.

Strain gages currently suffer from two important limitations which polymeric strain gages should alleviate. (1) They require bulky and sensitive external electronics to amplify the signals generated by the strains and (2) dynamic range is limited at the upper end by the 0.1 % recoverable strain. The latter range could be extended by the use of polymers, in which recoverable strains of between >1 % (glassy) and >100 % (elastomeric) may be

achieved. Until recently, however, polymers have been ruled out due to their highly insulating nature. Furthermore, as will be described, it will be possible to integrate polymer-based electronics and energy sources with polymer strain gages, forming compact, materially compatible and cost-effective devices.

Experiment

Polypyrrole films are employed in the strain gage tests. The mechanical electrical and chemical properties of these films have been described in great detail in previous reports. The polypyrrole is galvanostatically polymerized on a glassy carbon substrate from mixture of 0.06 M freshly distilled pyrrole monomer and 0.05 M tetra ethyl ammonium hexafluorophosphate in propylene carbonate. Deposition takes place at $-30\ ^oC$ in a nitrogen atmosphere at a current density of 1.25 $A\ m^{-2}$. Film dimensions are typically 90 mm long x 6 mm wide x 40-100 μm thick. The resulting material exhibits a density of 1.4 $\times 10^3\ kg{\cdot}m^{-3}$, a conductivity between 1-3x10^4 $S{\cdot}m^{-1}$, a glassy modulus of 0.5 GPa (wet) and 1.0 GPa (dry), and a tensile strength of > 25 MPa. Some films were then stretched uniaxially by 30 % in a propylene carbonate bath at 40 °C.

Figure 2: Gage Factor test set-up
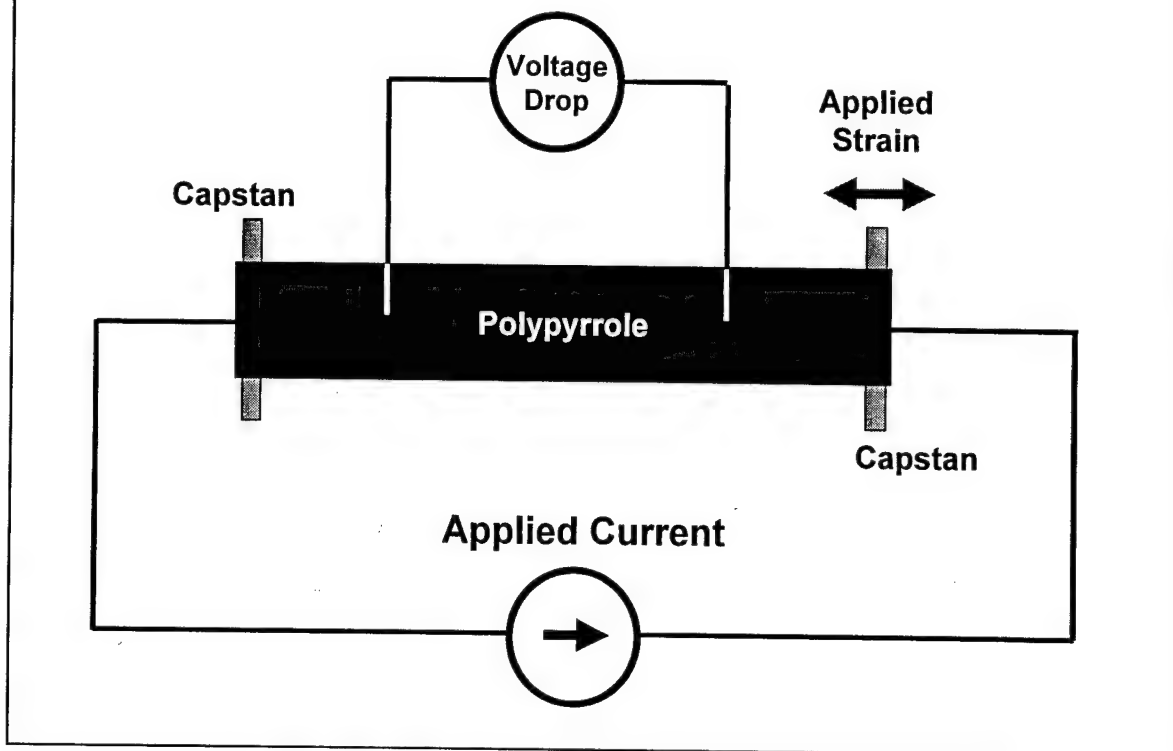
Four point resistance measurements are made on the polypyrrole films while applied strain and stress are ramped. Figure 2 is a schematic of the basic test set-up, and Figure 3 is a photo of the apparatus. The films ends are wound around 2.5 mm diameter slotted rods, providing mechanical coupling to the rods via the capstan effect. The rods are displaced using galvanometers (General Scanning, Cambridge MA, Model 350). Film displacement is measured using angular displacement transducers built into the galvanometers, while force is proportional to the applied motor current.
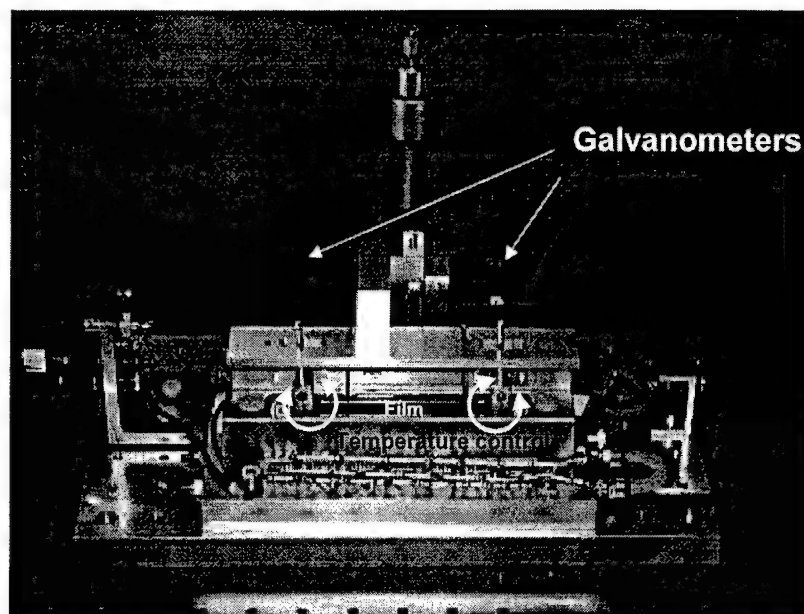
Figure 3: Photograph of gage factor test apparatus.

In order to measure resistance, a 15 *mA* current is applied at either end of the film. Gold wire contacts are attached to the film at a 30 mm spacing using conductive carbon adhesive (Electron Microscopy Sciences, Pennsylvania). The product of the voltage drop between the gold wire contacts and the applied current is the film resistance. Given the distance between contacts and the film cross-sectional area, film resistivity is determined. Resistivity is typically in the range of 0.5 and 1×10^{-4} ohm·meters, between that of silicon and mercury (Figure 4). Forces or displacements are applied to the films and the resulting changes in resistance recorded.

Figure 4: Resistivity of polypyrrole relative to metals and semiconductors.
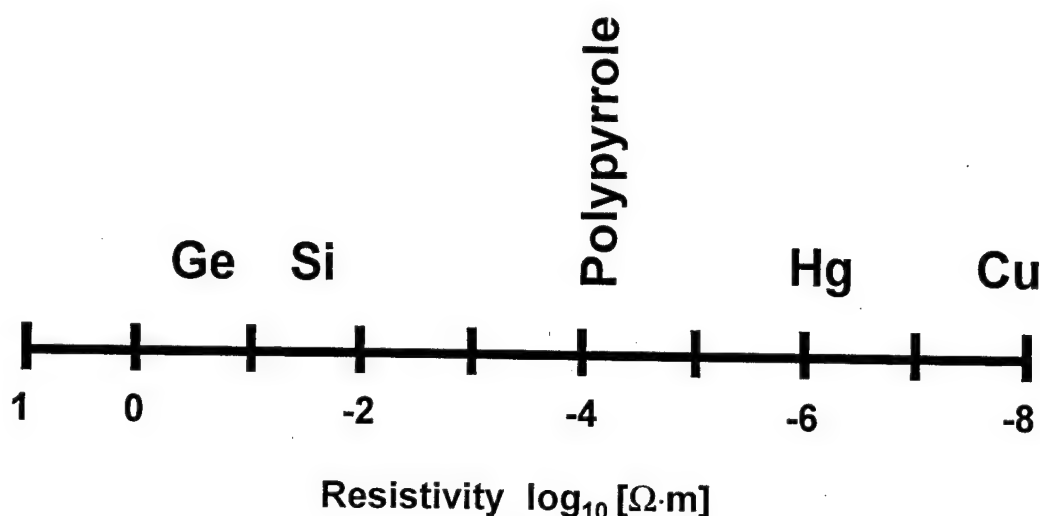
Results: Gage Factor

Figure 5 shows the relative change in resistance of an unstretched polypyrrole film as a function of strain, in response to a 0.005 Hz strain rate. The corresponding gage factor is 0.18. Note the linearity of response despite the large deformation. Further experiments were performed at strain rates between 0.000005 and 0.01 Hz, which demonstrated that gage factor is independent of rate. The results are promising because they demonstrates that static linear response can be obtained from polypyrrole strain gages over a strain range of at least 1 %, ten times the range of traditional gages. Change in resistance is also related to film stress, as shown in Figure 6. The relationship between relative resistance and stress is linear over the measurement range, with a slope of 3.4×10^{-4} MPa^{-1}.

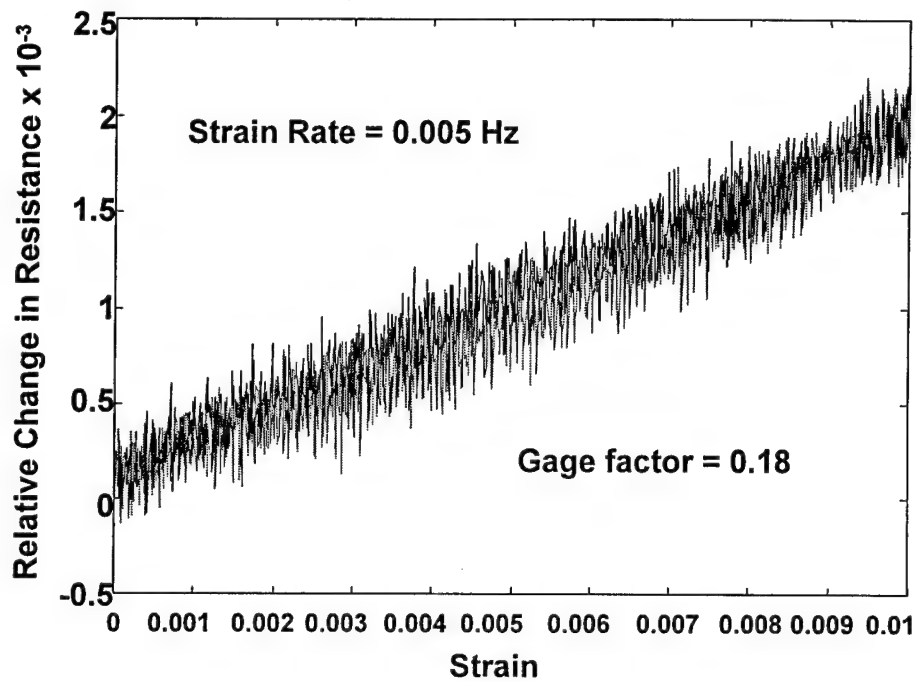Figure 5: Gage factor determination in isotropic polypyrrole.

Strain Rate = 0.005 Hz

Gage factor = 0.18

Figure 6: Relationship between film stress and relative resistance change in isotropic polypyrrole.

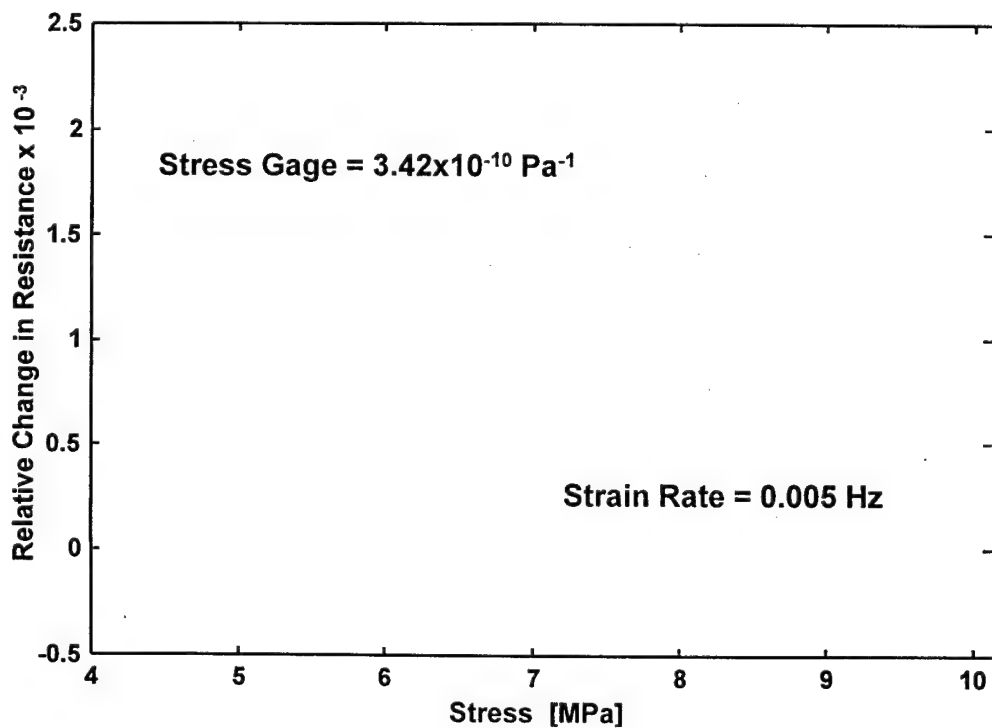Stress Gage = 3.42×10^{-10} Pa^{-1}

Strain Rate = 0.005 Hz

Figure 7: Creep in an isotropic polypyrrole film.

5 MPa
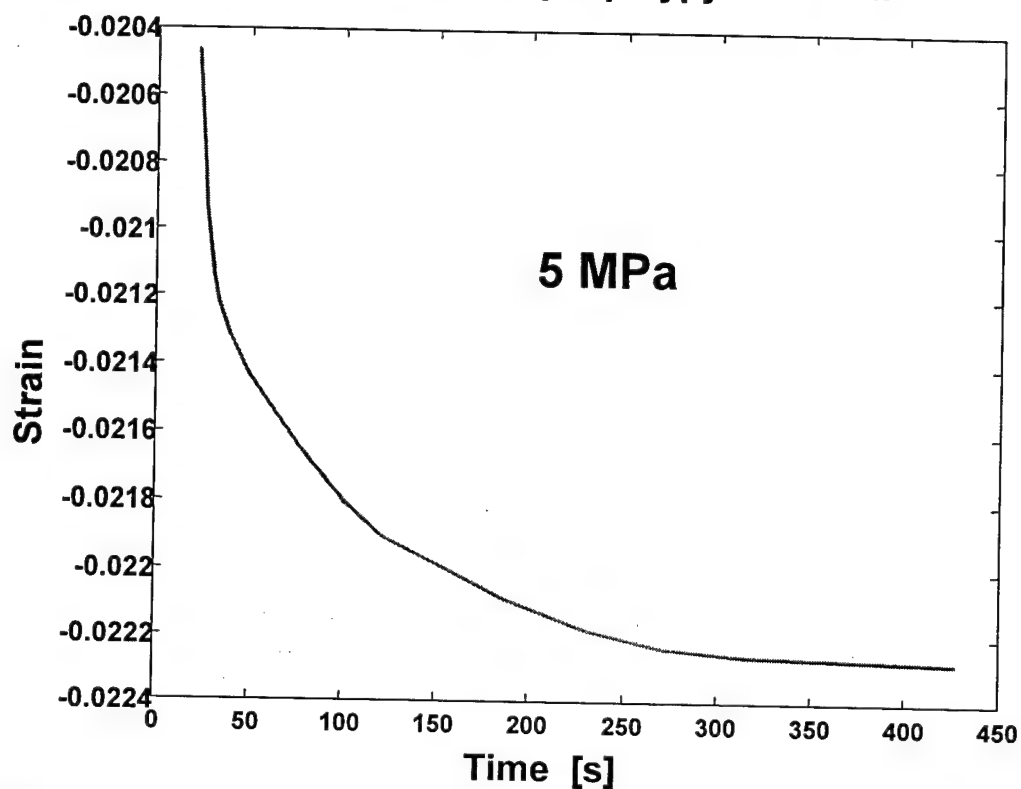
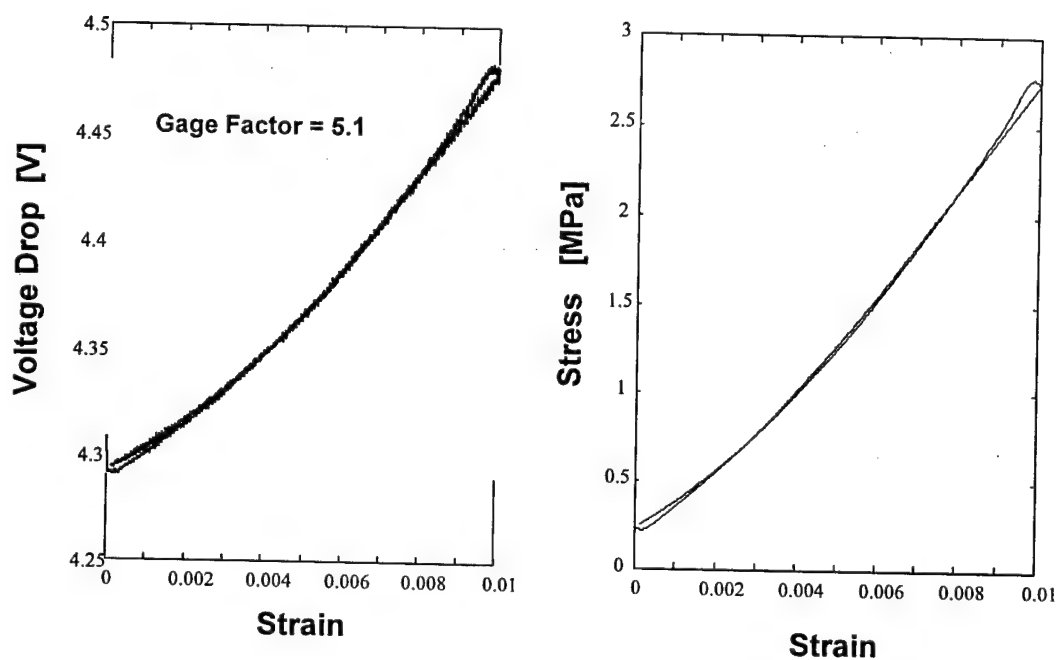Figure 8: Gage factor and stress-strain curve of stretch aligned polypyrrole

Gage Factor = 5.1

Figure 9: Table of gage properties.

Properties	Constantan	P-doped Si	Polypyrrole
Gage Factor	2.1	40-200	0.2-5
Strain [%]	0.1	0.1	>1
Temperature Coefficient [K^{-1}]	20	300-4000	300
Resistivity [$\Omega \cdot m$]	1.7×10^{-6}	3.8×10^{-2}	1×10^{-4}
Temperature Range [$^{\circ}C$]	-195 to 315	-	< 110

The forthcoming research phase will involve the testing of gage performance as a function of cycle number, temperature and degree of stretching, followed by integration with polymer transistors and batteries. However, polypyrrole-based strain gages already compare favorably with traditional transducers, as outlined in Figure 9. The integration of polymer transistors with the strain gages will be discussed after the following overview of electrochemical device characterization.

Electrochemistry of Polypyrrole

Conducting polymer based transistors, batteries, electrochromic devices and actuators all rely on material properties that are functions of polymer oxidation state, and therefore are electrochemically tunable. As oxidation state changes, electrochromic devices shift

11

absorption bands, transistors change in conductivity and batteries charge or discharge. It is important to determine the range of potentials over which such transitions occur. If the range is exceeded, then energy may be dissipated unnecessarily, and harmful side reactions may be induced. The cell potential determines the energy density of batteries, the efficiency of actuators and of electrochromic devices, and the transconductance of transistors. In this section, we demonstrate methods of experimentally determining the potential range.

An electrochemical cell consists of two electrodes, each in contact with reactants and products, and between which is an ionically conductive medium (but not electronically conductive) known as the electrolyte. Each electrode has a half reaction associated with it, one involving oxidation (anode), and the other reduction (cathode). The electrodes are in electrical contact, with a potential difference being applied or generated between them and a current flowing through the circuit.

Each half reaction may be driven forward or in reverse, depending on the cell potential. An equilibrium potential exists at which no current flows. The equilibrium potential is determined by the relative concentrations of reactants and products at each electrode, and is purely a thermodynamic quantity. Half reaction reduction potentials are thermodynamic constants whose values have, in many cases, been tabulated relative to certain standard and well characterized electrodes (i.e. the Normal Hydrogen Electrode, whose potential is defined as 0 V). When the cell potential deviates from equilibrium, current flows, the amount of charge transferred being proportional to the extent of reaction.

The magnitude of the current is generally limited by one of two factors, either the rate at which reactants can reach the electrode, or the rate of reaction. The former case is referred to as a mass transport limited or reversible reaction and the latter a kinetics limited or irreversible reaction. In a reversible reaction, since the kinetics are relatively fast, a new equilibrium is reached at the electrode surface, as determined by the applied potential, which sets the equilibrium concentration of reactants and products. These concentrations are different from those in the bulk material, inducing mass transport due to concentration gradients (diffusion). The rate of diffusion is proportional to the difference in concentration between the electrode surface and the bulk electrolyte. Therefore, as potential deviates increasingly from the equilibrium, diffusion becomes increasingly rapid. The extent of this deviation is known as the overpotential. The rate saturates at overpotentials beyond which the concentration of reactants at the surface is essentially zero, the saturation overpotential typically being less than 120 mV. In irreversible processes, kinetics are limited by an activation barrier. Increasing the magnitude of the overpotential increases reaction rate exponentially until mass transport becomes the limiting factor. In both the reversible and irreversible cases it is useful to know the shape of the I-V curve as this determines tradeoff between rate of change of oxidation state and applied potential. In the subsequent discussion we demonstrate how to measure the I-V curve. The V/I transfer function for a reaction at an electrode is referred to as the faradic impedance.

Another source of potential drop occurs when current begins to flow in a cell. By Kirchoff's law, current must be equal in all parts of the circuit. Thus a current must flow through the electrolyte. Because the electrolyte is not electronically conductive, charge

transport is due to migration of ions. These ions encounter viscous drag forces as they move through the solution, and therefore have an associated resistive drop. The magnitude of this drop is a function of the conductivity, and can be substantial, especially in cases of high current densities, low ionic concentrations and large electrode spacings. We show in the following section how to measure solution conductivity, thereby enabling the magnitude of potential drop due to solution resistance to be calculated for a given cell geometry and current. In general one seeks to minimize the solution potential drop in a device as it results in energy dissipation.

Before proceeding to describe the measurement of the electrode and solution impedance characteristics, it is important to discuss electrode capacitance that results from the build-up of ionic charge at an electrode surface. Imagine a non-reacting electrolyte solution containing, for example, 1 M NaCl in water. Application of a 1 V potential difference between two inert electrodes will, at first, generate a field gradient in the solution. This gradient causes the chlorine ions to move towards one electrode and the sodium ions to travel in the opposite direction. Since no reaction occurs at the electrodes at this potential, ions very rapidly accumulate next to the electrodes. The charging is balanced by the accumulation of electronic charge at the electrode surface. The charged regions at the electrodes are referred to as double layers, and act as capacitors, canceling the field across the bulk solution. The time constant associated with this charging is a function of the capacitance at the electrode interfaces and the solution resistance, and can be roughly modeled as a series RC circuit. If ac potentials are applied at frequencies that are large relative to the inverse of the RC time constant, the capacitance becomes negligible relative to the solution resistance, enabling the latter to be measured.

14

Figure 10: Test cell for measuring solution conductivity.
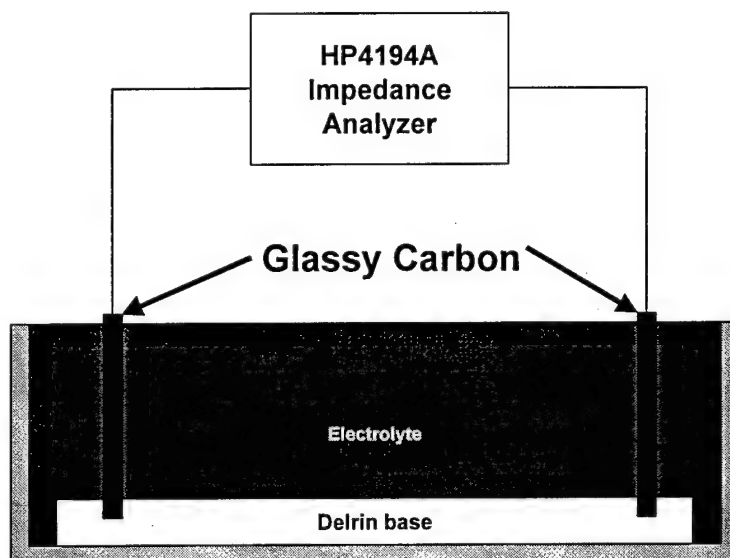
The double layer is also present when electron transfer is occurring at the electrode:electrolyte interface. It usually provides the potential difference that drives a reaction. Thus the double layer impedance and the Faradic impedance act in parallel. At high applied frequencies, typically on the order of 1 to 30 kHz, the impedance at the interface becomes negligible compared to solution resistance.

Conducting polymers are unusual in that they act simultaneously as reactants, products and electrodes. The double layer then is no longer responsible for generating the potential drop that drives the Faradic reaction. Instead it is the potential difference applied between the conducting polymer film and the counter electrode that directly drives the reaction.

Solution Conductivity

The impedance of an electrochemical cell has two components: the electrolyte resistivity and the potential drop at the electrode surfaces. The former is the result of the energy required to propel ions through a viscous medium, and is a major cause of internal resistance. We seek to determine the ionic conductivities of electrolytes such that their internal resistances can be predicted for a given cell configuration.

Given the capacitive nature of the electrodes, their impedance becomes negligible at frequencies above about 1 kHz, allowing the component due to ionic resistivity to be extracted. The results shown in Figure 11 are typical of electrolyte impedance as a function of frequency. The top plot shows magnitude and the bottom phase. Initially, phase is negative, indicating a capacitive component to the response. However, at 1 kHz, the phase is nearly zero and the magnitude has decreased, indicating that the capacitive component has diminished. Between 1 kHz and 100 kHz the magnitude of impedance is nearly constant and the phase is zero, as is characteristic of a purely resistive response. It is from this frequency region that measurements of solution resistance are taken.

Resistance is an extrinsic quantity that depends on the cell geometry. In order to obtain resistivity and conductivity, either the geometry must be accurately known and modeled, or the resistance of a reference solution of known resistivity, is required. In the latter approach, the ratio of measured resistances is equal to the ratio of resistivities, enabling the unknown resistivity to be determined.

Figure 11: Impedance of 0.05 M tetraethylammonium hexafluorophosphate.
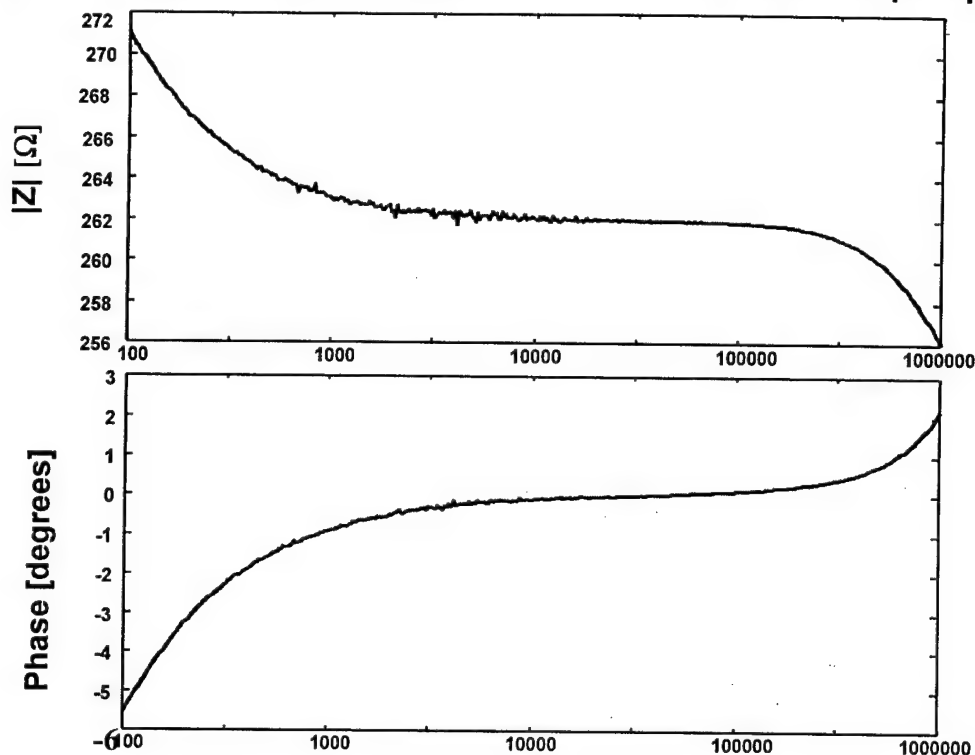
Experiment

The experimental apparatus employed for measuring electrolyte is depicted in Figure 10. The electrodes are 50 mm square, 2 mm thick glassy carbon plates (Alpha-Aesar, Wilmington MA). These are mounted in slots machined in a delrin base such that the electrode separation is 140 mm. The electrodes and base sit in a 160 mm long, 100 mm wide and 70 mm deep glass bath which is filled with electrolyte such that the top 5 mm of the electrodes protrude from the solution. Wires are attached to the protruding electrodes and connected to an HP 4194A impedance analyzer. Impedance magnitude and phase are measured using a swept sine of 0.5 V amplitude at 401 equally spaced frequencies on a logarithmic scale, between 100 Hz and 1 MHz.

Figure 12: Impedance of 1.4 mM tetraethylammonium hexafluorophosphate
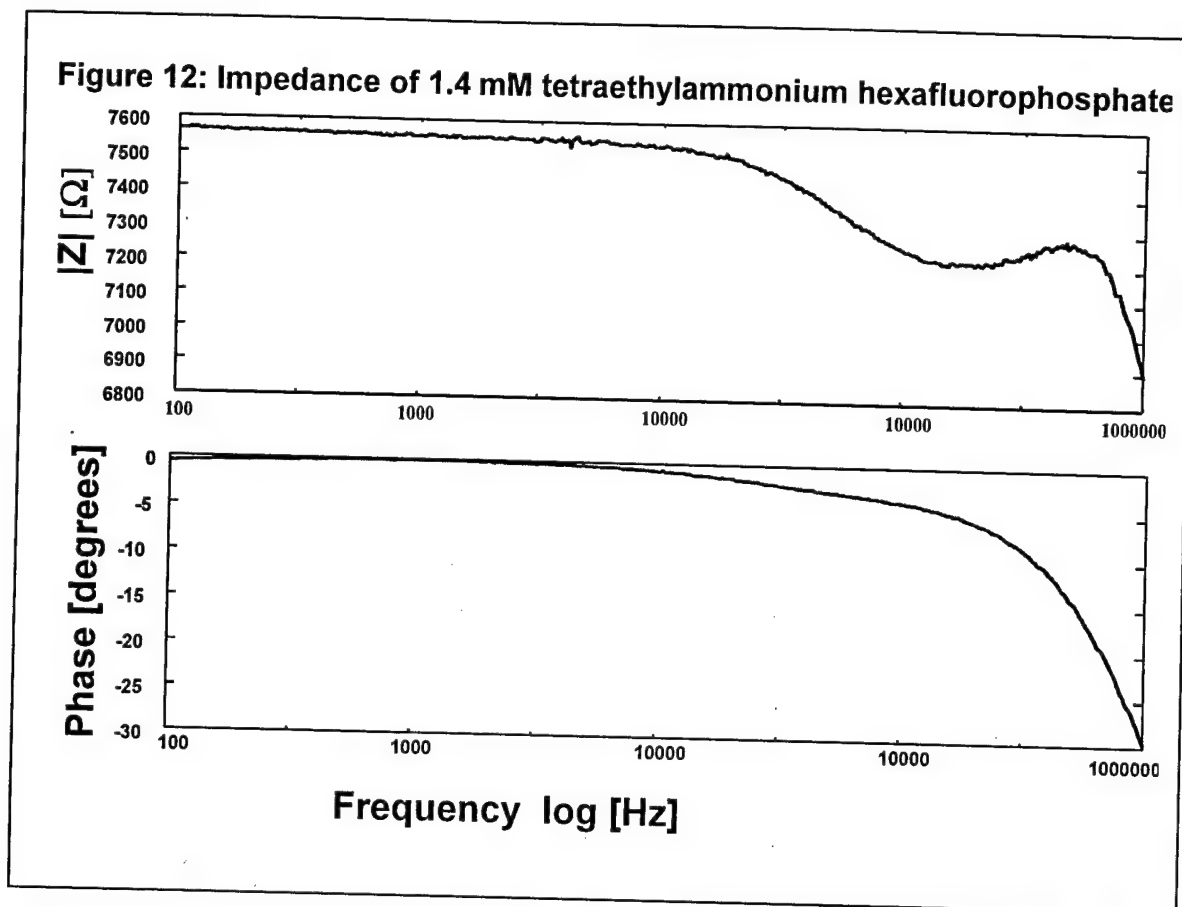
The impedances of two solutions are measured, 0.05 M tetraethylammonium hexafluorophosphate in propylene carbonate, and 0.0014 M tetraethylammonium hexafluorophosphate in propylene carbonate. The second solution acts as a reference, having a known conductivity of 0.00432 $S \cdot m^{-1}$.

Results

Figures 11 and 12 display the measured frequency responses for the two solutions. In Figure 11 the phase is less than 5 m° at 30 kHz, and the corresponding magnitude is 262 Ω. In Figure 12, it is clear that the capacitive effect at low frequencies is relatively small. This is because of the low ionic concentration, which in turn reduces the charge density at the electrodes, forming instead a fairly diffuse double layer. Resistance is again

measured at the frequency at which phase is closest to zero, which occurs at 1.22 kHz, with the corresponding magnitude of impedance equal to 1550 Ω. Both measurements were taken at 24 °C. The conductivity of the 0.05 M solution is thus 0.12 S·m^{-1}.

I-V Curve

In the previous section it is demonstrated that electrolyte resistance dominates cell impedance at high frequencies. Most of the effects of interest occurring in a polymer, however, result from change in oxidation state, and these are observed at low frequencies. For example, color and conductivity are both functions of oxidation state. Here we seek to determine at what applied potentials these Faradic processes take place.

As mentioned above, both Faradic and capacitive currents are generated at low frequencies. Since it is the former that generates the changes in material properties of interest, these must be distinguished from the capacitive currents. One method is to employ chronoamperometry. In this experimental procedure, step changes in potential are applied across an electrode and the resulting current is recorded as a function of time. Any current resulting from double layer charging shows a characteristic exponential decay in response to a stepped potential. Faradic current, on the other hand, will typically exhibit a power law time dependence, which will be apparent even once the capacitive current has become negligible. Such behavior is observed in Figure 15. A one volt step is applied to a polypyrrole coated electrode, resulting in the current shown. The current decays exponentially at first, but then flattens, as the faradic current becomes dominant. By sampling current at a time, T, after the application of the step change in applied potential, and plotting the recorded current as a function of step magnitude, it is possible

19

to determine the potential dependence of faradic current. This works providing the time, T, is chosen such that the capacitive current is negligible relative to the faradic contribution.

A cell is composed of two half reactions. If a step change in potential is applied across the cell, then each electrode will account for a portion of the drop. However, we are interested in the behavior at one electrode and need to distinguish the potential drop across it from that at the counter electrode. This requires the use of a third electrode, known as a reference electrode. The third electrode is placed next to the electrode of interest (the working electrode), in order to minimize any potential drop due to solution resistance. The reference electrode is chosen to have a well-established and stable reduction potential. In order to maintain a stable potential, the concentrations of reactants and products must remain constant. The reference must have a high impedance to minimize current, thereby eliminating over-potential.
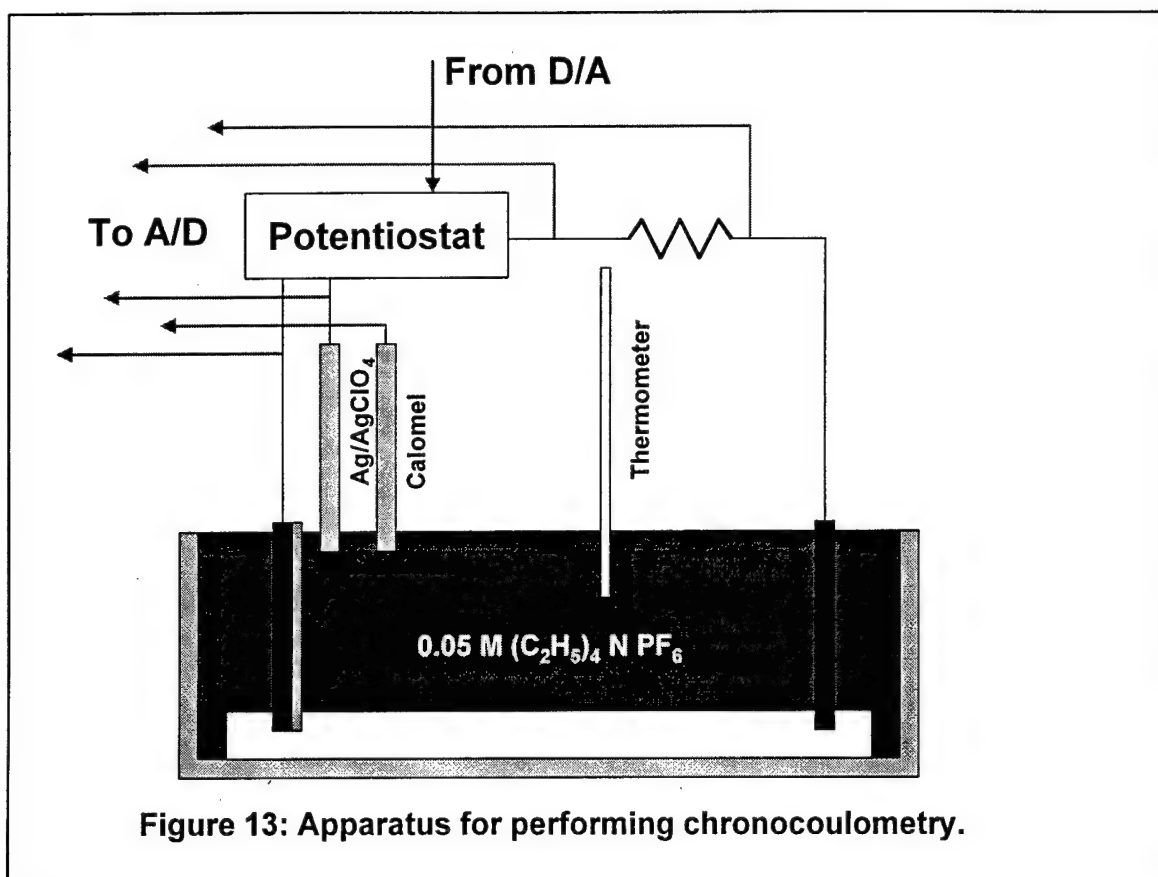
**Figure 13: Apparatus for performing chronocoulometry.**

For chronoamperometry to work effectively, not only must the potential across the working electrode be measured, but it must also be set. If a step potential is simply applied across the cell, the potential drop will be divided between the two electrodes and the electrolyte. The relative magnitudes of these are likely to be time dependent, and any change in potential across the working electrode generates further capacitive current, counter to the aim of minimizing this current. In order to control the potential drop across the working electrode, a feedback loop is required which adjusts the cell current such that the commanded working to reference electrode potential is maintained. An example of such a circuit, known as a potentiostat, is shown in Figure 14. The command potential sets the work to reference potential. A current buffer is employed to maximize

reference electrode impedance. Using the potentiostat, voltage becomes the independent

variable, and current is dependent.

Experiment and Results

The apparatus employed to perform chronoamperometry is shown in Figure 13, and the

circuit diagram is depicted in Figure 14. The same cell that is employed to perform the

electrolyte conductivity measurements is again employed here. The working electrode is

again a glassy carbon plate, but is now coated with a 10 micrometer thick layer of

polypyrrole electro-deposited as described above. Two reference electrodes are

Fig 15: Current in response to a -1 V step applied vs. a silver/silver perchlorate reference electrode for 20 s.
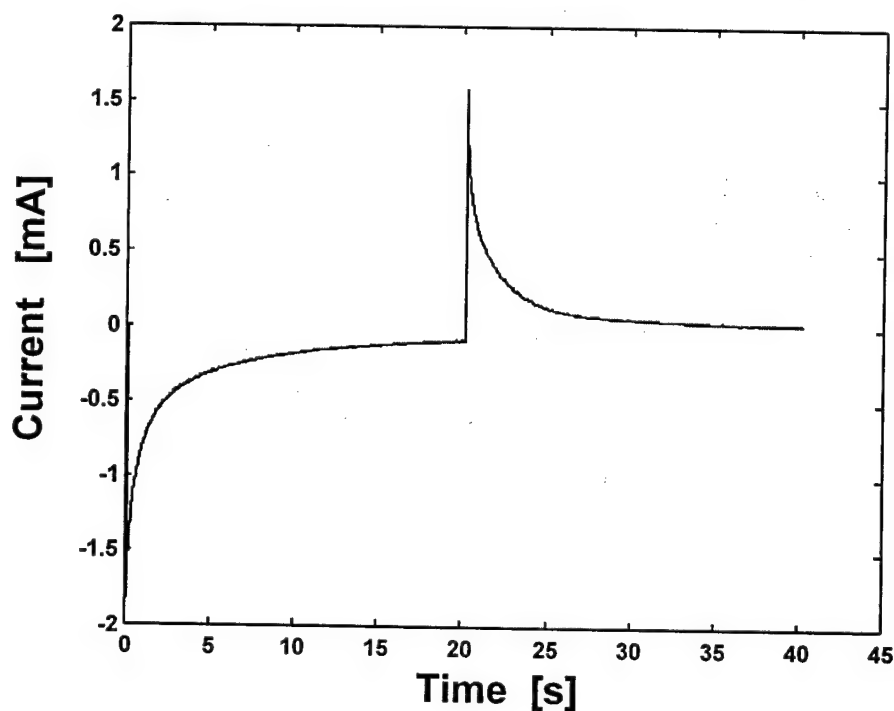
employed, one a calomel electrode and the other a silver/ silver perchlorate reference.

The later consists of a silver wire inside a glass tube that is capped with a viton ion-

exchange plug. The tube is full of 0.05 M tetraethylammonium perchlorate and 0.005 M

silver perchlorate in propylene carbonate. The half reaction is:

$$Ag\ (s) \leftrightarrow Ag^+(pc) + e^-$$
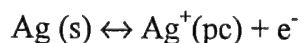
The silver perchlorate potential is employed in the potentiostat circuit, with the calomel

potential being used to check on the former electrode's stability. Otherwise the cell is

Figure 16: Step potential voltammetry results from a polypyrrole film in 0.05 M tetraethylammonium hexaflurophosphate
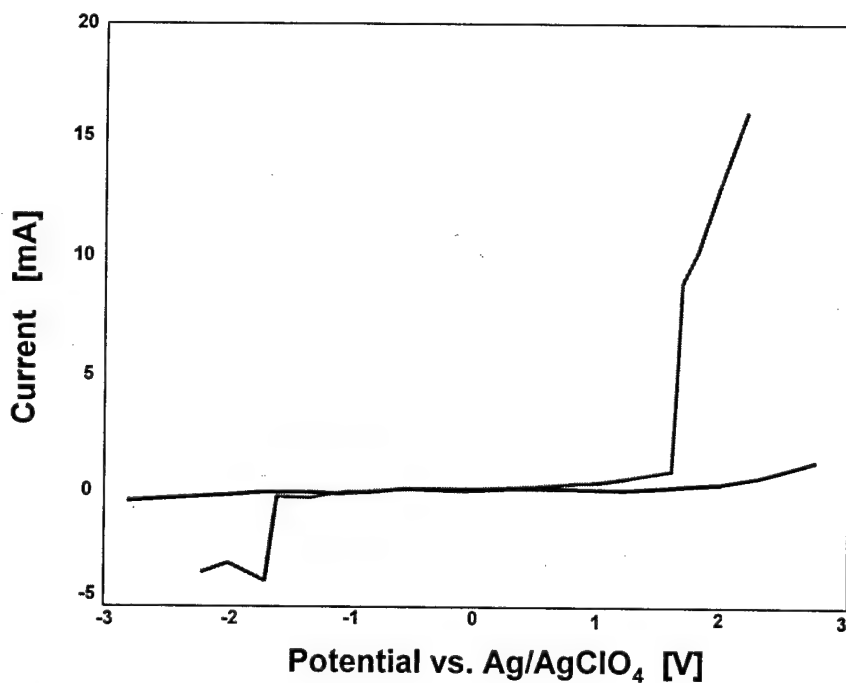
identical to that used in measuring conductivity.

The experimental procedure begins with the application of a step potential between the

working and the reference electrodes. This potential is held for 20 s, and then removed

for a further 20 s. During this 40 s period, current, cell potential and work-reference

potential are all recorded. The magnitude of the step is increased from −1.2 V to 1.2 V in

23

0.2 V increments. A typical record of current versus time is shown in Figure 15, where the current is in response to a −1 V step. The current measured at 20 s after the application of the step is then plotted as a function of the step magnitude, Figure 16. The same procedure was applied to cell having a working electrode without pyrrole to determine the extent of the background current due to possible electrochemical reaction of the electolyte or solvent. This background current is significantly smaller than that observed when the polypyrrole is present, as seen in Figure 16.

The results of step voltammetry shown in Figure 16 indicate the onset of oxidation at 1.2 V vs. Ag/ AgClO$_4$ and reduction at −0.8 V vs. Ag/AgClO$_4$. The I-V curve is now determined for polypyrrole in its as grown state.

Analysis of Polymer Transistor Gain

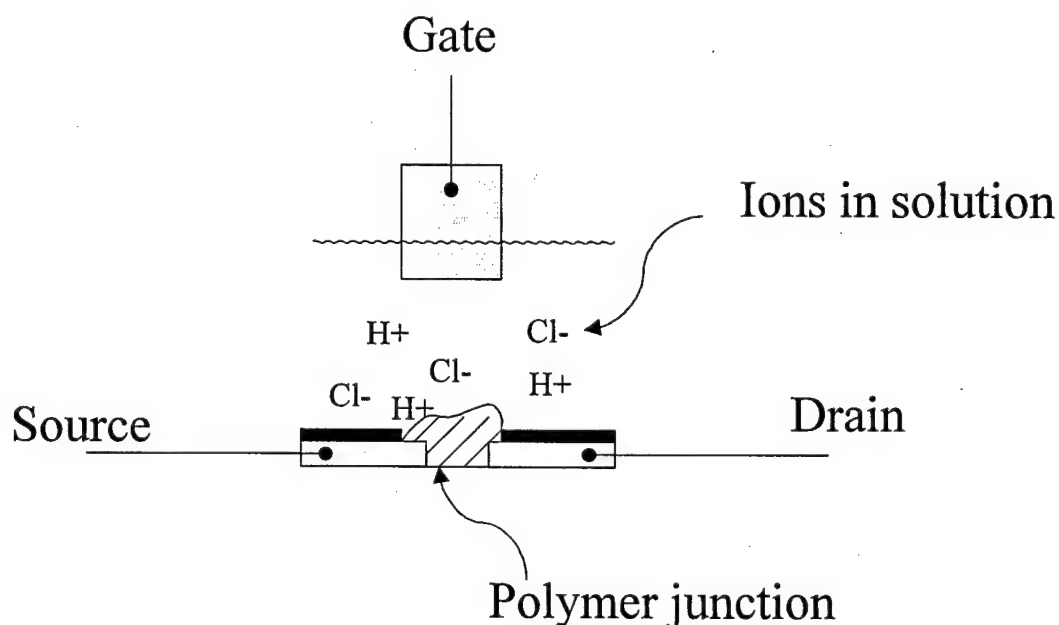The voltage signal from polymeric strain gage (or any other polymer sensor device) may

Figure 1: Polymer transistor diagram. In operation, a small voltage bias is applied between the source and drain. The resistance of the polymer junction is changed by driving Cl ions into and out of the polymer, thereby changing

need to be amplified to drive following circuits. To simplify construction of a complete integrated device, we would like to amplify the signal using a polymer transistor. We are currently investigating a polymer transistor amplifier where the gate voltage of the polymer transistor controls the transistor source to drain resistance.

A diagram of a typical polymer transistor is shown in Figure 1. The transistor operates in an electrochemical solution. Chlorine ions in the solution are driven into and out of the polymer junction between the source and drain. The ions induce a change in the conducting state of the polymer, which changes the source to drain resistance.

Polymer transistor switching characteristics for a gold gate transistor are shown in Figure 2. As the gate voltage increases from 0.1V to about 0.4 or 0.5V, the transistor switches from its on state (low resistance) to its off state (high resistance). Note that if the gate material is changed, the voltage at which the switching occurs will also change.
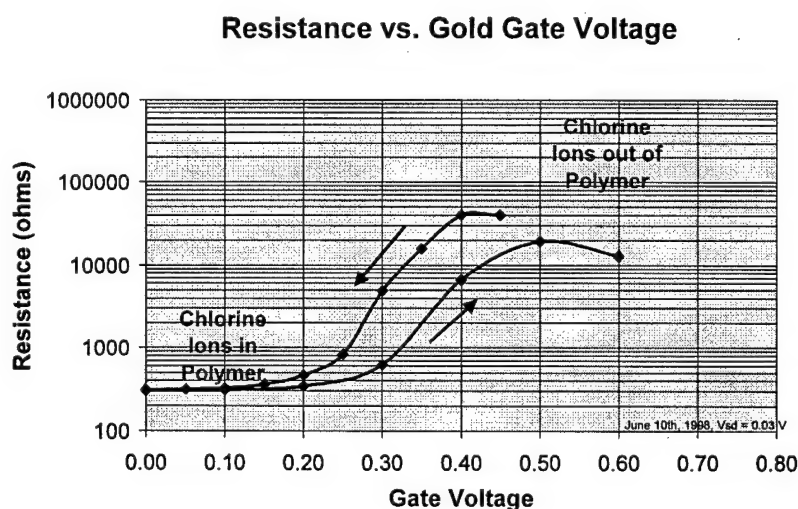
Resistance vs. Gold Gate Voltage

Figure 2: Plot of source to drain resistance vs. gate voltage for a PANI polymer transistor. (Note that the resistance measured between 0.5 and 0.6 V is constant within measurement error: at the high resistance state, the measurement error was as high as 100 to 200% because of errors in the current measurement). The source drain voltage was 50 mV.

25

In the on state, below a gate voltage of about 0.1 V, changes in the gate voltage no longer decrease the resistance. At this voltage the polymer material is saturated. Ions from the electrochemical solution have flowed into the polyaniline and created charged state carriers at all the available sites on the polymer backbone. Likewise, when the gate voltage reaches 0.4 to 0.5 V, the polymer reaches its maximum on voltage. At this voltage, the chlorine ions have mostly left the polymer junction.

In one mode of transistor operation, a small voltage is applied between the source and the drain. The transistor then acts as a current source, where the value of the current is dependent on the gate voltage:

$$i_{sd} = \frac{V_{sd}}{R(V_g)}$$

where i_{sd} is the source to drain current, V_{sd} is the source to drain voltage, V_g is the gate voltage, and R is the source to drain resistance and is a function of V_g. The plot in Figure 2 was obtained by measuring i_{sd} at constant source drain voltage.

A more useful configuration for building a voltage amplifier is shown in Figure 3. By connecting the source to drain resistance in series with a load resistor, the current through the load resistor (and hence the voltage across the load resistor) is a function of the source to drain junction resistance.
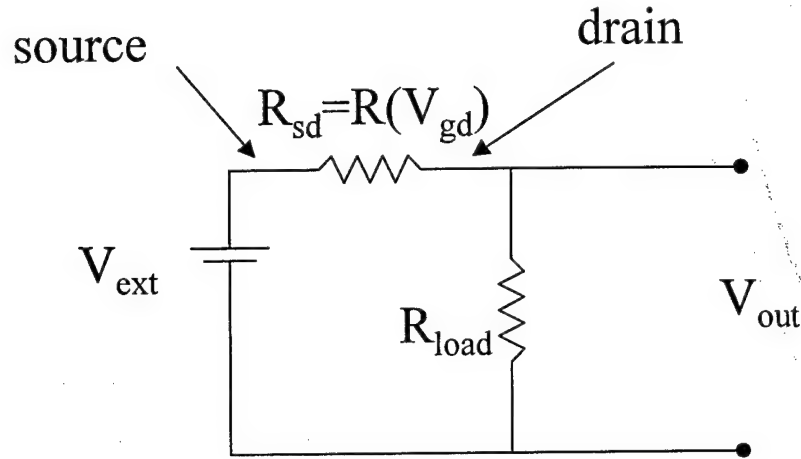
Figure 3: Diagram of voltage amplifier circuit. The source to drain voltage is connected in series with a load resistance to create a voltage dividing circuit. As the gate voltage is changd, the resistance R_{sd} changes, in turn causing the output voltage V_{out} to change.

In order to determine the voltage gain of the circuit, the relationship between source-drain resistance and gate voltage needs to be known. In our first generation amplifiers, we will operate the polymer transistor in the transition region between the on state and the off state. In this region, over a gate voltage range of about 100 mV the slope is close to linear (on a $\log(R_{sd})$ vs. gate voltage plot)[1] and so we write:

$$\log(R_{sd}) = mV_g + C$$

where m is the slope of the line and C is a constant. Now, let V_g be constant at some initial value and add a small voltage ΔV_g:

---

[1] There is significant hysteresis between the downward and upward curves in the measured data. However, in operation of the transistor, only small signals will be applied to the gate voltage and the large hysteresis observed due to slow ion diffusion will not be as important.

$$\log(R_{sd} + \Delta R_{sd}) - \log(R_{sd}) = m\Delta V_g$$

(ΔR_{sd} is the change in source drain resistance caused by the change in gate voltage). By

rearranging and taking the exponential, we find:

$$\frac{R_{sd} + \Delta R_{sd}}{R_{sd}} = \exp(m\Delta V_g)$$

which we can expand (for $m\Delta V_g \ll 1$):

$$1 + \frac{\Delta R_{sd}}{R_{sd}} = 1 + m\Delta V_g$$

$$\frac{\Delta R_{sd}}{R_{sd}} = m\Delta V_g$$

where second order and higher terms have been neglected. The amplifier circuit can be

treated as a voltage divider:

$$V_{out} + \Delta V_{out} = \frac{R_{load}}{R_{load} + R_{sd} + \Delta R_{sd}} V_{ext}$$

where V_{out} and V_{ext} are as shown in Figure 3. By expanding the denominator in a Taylor

series and substituting for ΔR_{sd} we find that for small changes in gate voltage,

$$\Delta V_{out} = -\frac{R_{load} V_{ext}}{R_{load} + R_{sd}} \left(\frac{R_{sd} m}{R_{load} + R_{sd}} \right) \Delta V_g$$

and for a given V_{ext} and m, the maximum gain will be obtained for $R_{load} = R_{sd}$ so:

$$\Delta V_{out} = -\frac{V_{ext} m}{4} \Delta V_g$$

Thus, in order to get voltage gain, the product $V_{ext} m$ must be maximised.

In the transistors built so far, we have obtained m values of about 10. For

electrochemical reasons, V_{ext} should not be increase much beyond 2V (for polyaniline).

Thus for transistors we have already built, we expect voltage gains on the order of 5. The

m value is a function of the polymer itself, the quality of the polymer deposition and

possibly of the junction geometry. We expect to be able to improve *m* by as much as 6 to 10 times, thereby obtaining gains as high as 50.

While gains of 50 are not high when compared to the gains obtained using standard silicon, we believe that the advantages of co-fabrication (where the sensor and amplifier are built from the same material) will often outweigh the benefits of better performance obtained by mixing sensor and amplifier materials. If higher gain is required, cascaded polymer amplifiers can be used. We are also investigating alternative amplifier configurations that make use of the saturation regions of the polymer junction and that operate on completely different principles.

Conclusion

In this period of research we have achieved a number of significant milestones that relate to the ability to fabricate integrated and inexpensive polymer devices. The first polymer strain gage has been demonstrated, which achieves a gage factor of 5, combined with a recoverable strain of greater than 1 %. Polyaniline transistors are demonstrated, and the feasibility of their use in a miniaturized strain gage/amplifier package analyzed. Finally, methods of electrochemical device characterization are demonstrated which are essential to understanding device behavior and determining efficiency.

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Sensors and Actuators

CHAPTER 11

Progress Towards a New Class of Photon Force-Based Chemical Sensor
C.J.H. Brenan, I.W. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Progress towards a new class of photon force-based chemical sensor

Colin J.H. Brenan and Ian W. Hunter
Department of Mechanical Engineering, Room 3-147
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

1.0 Introduction

Detection of biological and chemical agents is of increasing importance for home environment monitoring. Inexpensive environmental sensors and sensor networks as part of a feedback loop can maintain the home environment to within a comfort range selected by the home's occupants or determined from the physiological state of the occupants (a "smart" house). Viable sensor technologies for home environmental monitoring must have the requisite sensitivity, specificity and response time but, of equal importance, they must be inexpensive to manufacture since the sensor purchase price will ultimately determine if the sensor technology will find its way into the home. Thus, any viable home sensor technology must be cheap without sacrificing desirable sensor characteristics.

Consideration of most chemical and biological sensor technologies having the requisite sensitivity and chemical specificity have shown they are simply to expensive at present to penetrate the home sensor market. Spectrochemical analysis via various spectroscopies (e.g. optical/IR absorption, Raman scattering, nuclear magnetic resonance) have the sensitivity, accuracy and selectivity but the system components, despite recent advances in spectrometer miniaturization, optical fiber and photosensor detection technologies, still make their cost prohibitive to the home market. Electrochemical sensor technologies, such as carbon monoxide sensors, are at an advantage because they exploit the well-established mass production technologies for the manufacturing of electrical circuitry.

2.0 Photon momentum approach

Our novel approach is a potentially new path towards creation of cheap, sensitive, robust and highly selective biochemical sensors for the home market. We exploit the mechanical dynamics of a micrometer-sized particle held in a three-dimensional force field produced by a focused laser beam. The particle is functionalized to be sensitive to a specific class of biological or chemical agent and the presence of a particular material is detected through measurement of small changes in particle mass in response to a known applied force. The concept is somewhat analogous to surface acoustic wave (SAW) chemical sensors except for one important difference; namely, the photon force (PF) chemical sensor can potentially detect compounds in either liquid or air (gas) whereas the SAW devices (and most electrochemical sensors) are confined to detection of air borne pathogens or chemicals.
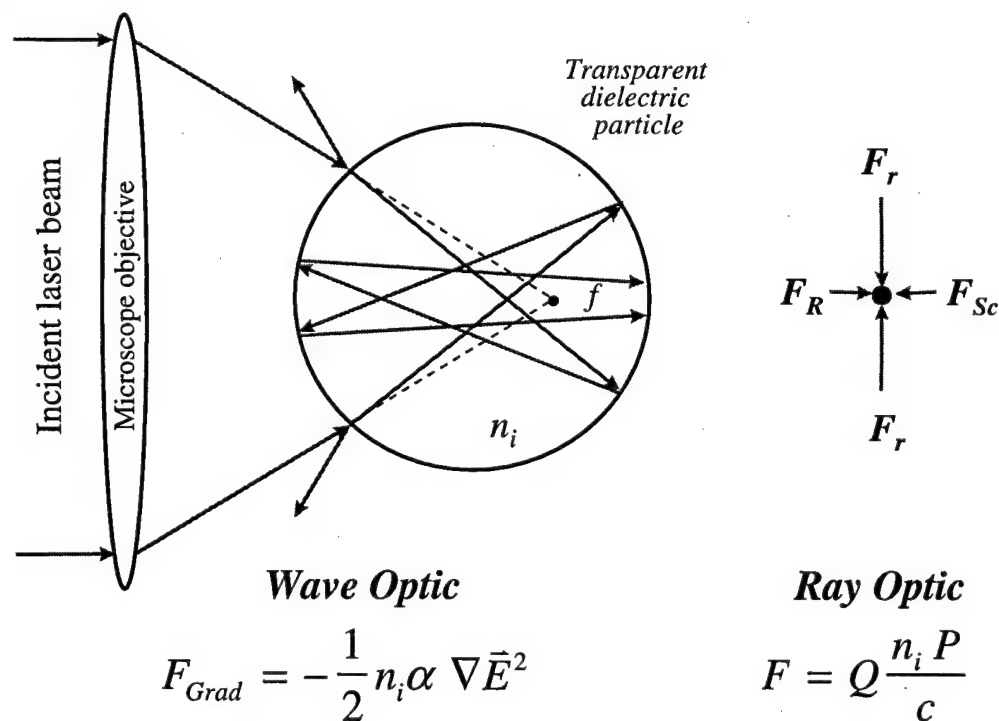

Wave Optic

$$F_{Grad} = -\frac{1}{2} n_i \alpha \, \nabla \bar{E}^2$$

Ray Optic

$$F = Q \frac{n_i P}{c}$$

Figure 1: Geometric optical description of photon force trap applied to a microscopic particle.

The idea at the heart of the PF sensor is diagrammed in Figure 1. Reflection of photons in a focused optical (laser) beam from the dielectric interfaces between the particle and its surroundings (either a gas or liquid) transfer linear momentum to the particle. If the focused light rays are incident on the particle at a large angle, axially symmetric radial and axial forces on the particle keep it trapped in a stable equilibrium close to the laser beam's geometric focus. In this position, the particle experiences a spring-like restoring force proportional to its displacement from the equilibrium point. The spring constant or stiffness of the optical trap is proportional to the optical power incident on the particle.
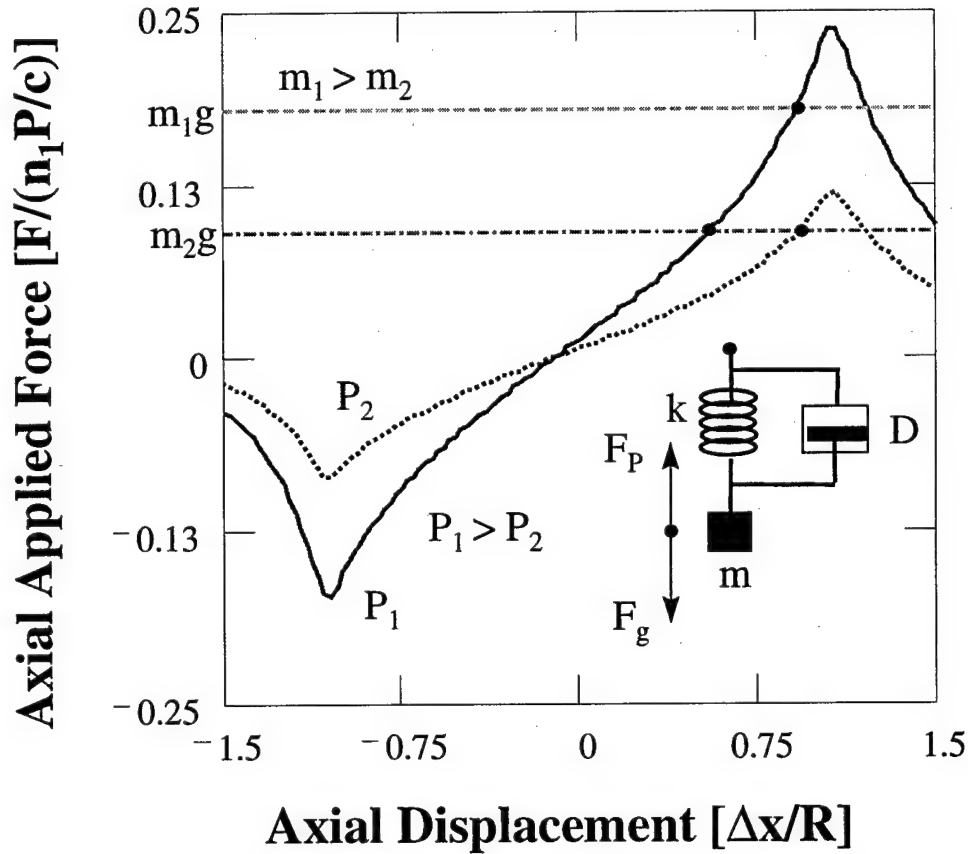

Figure 2: Result of calculation (in geometric optic approximation) showing axial trapping force as a function of axial displacement. Decreasing the optical beam power (P_1 -> P_2) decreases the trap stiffness and moves the particle to a new equilibrium position away from the geometric focus. The mechanical model for this system (see inset) is a damped harmonic oscillator where the damping force is derived from the viscous Stokes force as the particle moves through the surrounding medium.

3

For a vertically-oriented trap in an optically homogenous and isotropic environment, the particle sits at an equilibrium position determined by a force balance between the photon trapping (radiation) force and gravity. Decreasing or increasing the optical power in the trapping beam displaces the particle to a new equilibrium position (Figure 2). Similarly, as the particle is displaced from its equilibrium position by moving the trapping beam relative to the particle, a proportional restoring force acts on the particle to move it back towards the equilibrium position.

The optical force acting on the particle, therefore, can be changed by either displacing the particle relative to the focused optical beam or by amplitude modulation of the trapping beam optical power. Measurement of the particle position within the trap combined with knowledge of trap stiffness (equivalent spring constant) enables the applied force acting on the particle to be estimated.

A summation of the radiation, gravitational and viscous drag (Stokes) forces on the particle results in derivation of the differential equation governing particle motion with time,

$$\frac{d^2z}{dt^2} + \frac{\gamma}{2m}\frac{dz}{dt} + \frac{k}{m}z = \frac{F(t)}{m},$$

1

where m is the particle mass, k is the trap spring constant, $F(t)$ is the time-varying force applied to the particle and

$$\gamma = 6\pi\eta R$$

2

is the drag coefficient of the viscous (Stokes) force acting on a particle of radius R immersed in a medium with viscosity η.

There are different methods by which the particle mass is estimated from the dynamical model as expressed in Equation 1. A preferred approach is based on system identification concepts in which one expressly perturbs some parameter of a physical system and records the result so that a causal relationship can be established between a system's inputs and its outputs. Applying this generalized concept to the present

situation, there are two different, yet equivalent, analysis paths to follow. The first is to measure the particle displacement in time in response to a stepped force input. Fitting Equation 1 to the observed response allows the particle mass to be estimated since the other equation parameters are known beforehand. An equivalent approach is to work in the frequency domain and measure the mechanical compliance transfer function of the trapped particle. The input, in this case, is a time-varying force that can be a series of equi-amplitude sinusoidal force modulations at different frequencies or a known but stochastic force modulation. The subsequent particle displacement is then recorded and the compliance (output displacement/input force) transfer function, $G(\omega)$, estimated. The transfer function has magnitude, $|G(\omega)|$, and phase, $\Phi(\omega)$, components expressed as

$$|G(\omega)| = \left|\frac{z}{F}\right| = \frac{1/m}{\sqrt{\left(\omega_o^2 - \omega^2\right) + \gamma\omega/m}} \qquad 3$$

and

$$\Phi(\omega) = \tan^{-1}\left(\frac{\gamma\omega/m}{\omega_o^2 - \omega^2}\right) \qquad 4$$

where $\omega_o = \sqrt{k/m}$ is the trapped particle's natural oscillation frequency.

Although either the amplitude or phase part of the stiffness transfer function could be used to estimate particle mass, a phase transfer function measurement is preferred for two reasons. First, phase measurements are, in general, more precise than amplitude measurements thus enabling the measurement of very small mass changes. Second, significant changes in particle mass could result in correspondingly large changes in particle radius. If the technique to measure particle displacement is sensitive to particle size, as with light scattering, then the changing particle radius could unduly bias the mass measurement. A phase measurement rather than one based on light scatter amplitude would therefore be less sensitive to particle radius and minimize this potential source of inaccuracy.

Rewriting Equation 4 to express directly its dependence on particle mass results in a relationship of the form

$$\Phi(\omega) = \tan^{-1}\left(\frac{6\pi\eta R}{\sqrt[3]{4/3\pi\rho}}\frac{\omega}{\sqrt[3]{m^2\left(\omega_o^2 - \omega^2\right)}}\right) \qquad 5$$

where ρ is the particle density. Figure 3 plots the phase transfer function of glass spheres suspended in water and having different radii ($\rho_{glass} = 1500\ kg/m^3$, $\eta_{water} = 0.001\ kg/m\text{-}s$). Note the large difference in phase at any given frequency for different particle radii. Alternatively, the force modulation frequency could be kept constant and the phase angle between force input and displacement output monitored.
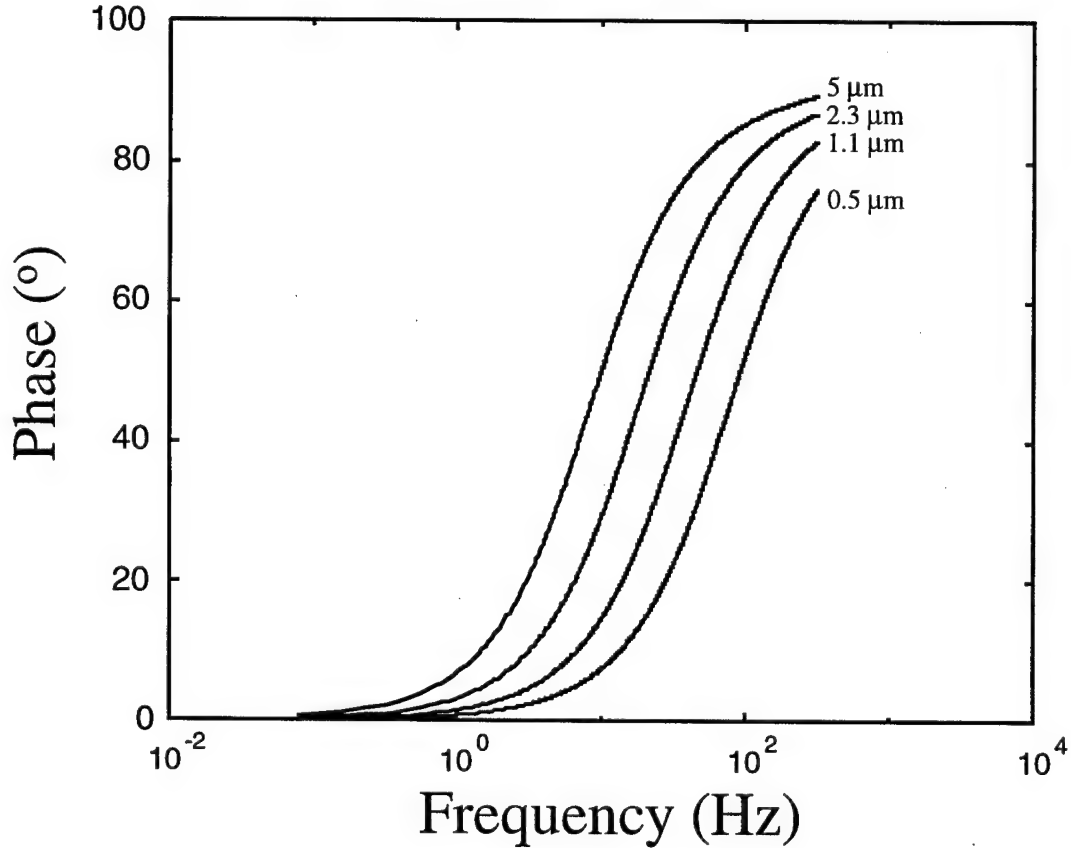

Figure 3: Theoretical phase transfer function of glass spheres in water for different particle radii.

Figure 4 shows the change in phase angle as a function of particle mass for different modulating frequencies. From this calculation it appears that the greatest sensitivity of phase to mass change occurs for high frequency perturbations applied to the particle. Phase angle measurement also allows the particle mass change to be measured over a large dynamic range, up to four orders of magnitude. The sensitivity of the measurement technique is expressed in Figure 5 where the change in mass for a given phase change is plotted as a function of particle mass for different modulation frequencies.
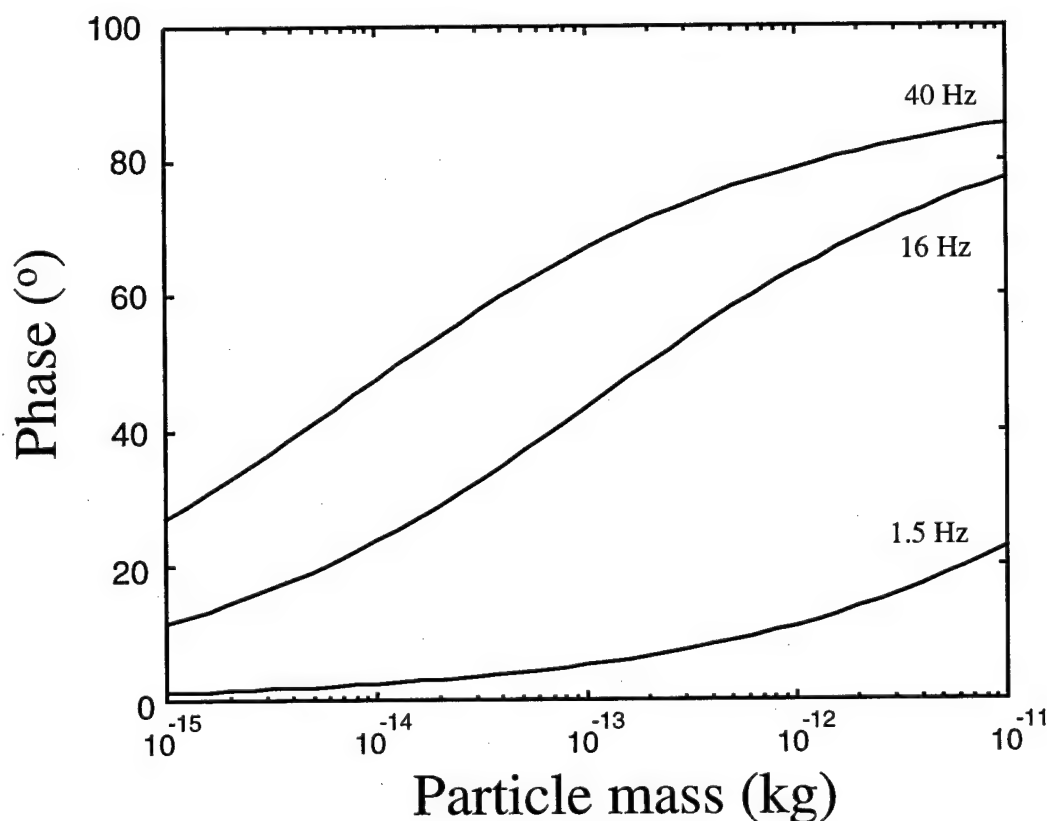

Figure 4: Phase angle as a function of particle mass for different force modulation frequencies.

What is the smallest mass change resolvable with this technique? To answer this question, the smallest resolvable phase angle, ϕ_{min}, is estimated assuming the largest source of error in determining the particle's position is from random fluctuations arising from Brownian motion. Assuming the particle is in thermal equilibrium with its

7

surroundings, the mean square particle displacement for one direction is $<\Delta x^2> = k_b T/k$ where k_b is Boltzmann's constant ($= 1.38 \times 10^{-23}$ J/K) and T is system temperature (300 ^oK). The particle displacement is measured by an optical system calibrated such that the particle's position is correlate with a measured photosensor current. A similar calibration is assumed to exist between the time-varying force applied to the particle (either by optical beam modulation or its relative displacement with respect to the particle). The two signals are recorded and then analyzed in one of two ways to find their relative phase difference.

The first method is to take the cross-correlation between the two signals, find the lag position relative to zero lag of the correlation maximum and calculate the phase difference between the two signals from the difference between the two lags. The second method is to multiply the two signals and low pass filter the output to get a DC power, i^2_s, proportional to $\sin \phi$. Averaged over N cycles of the sinusoidal force modulation the DC photocurrent is,

$$i_s = N\, R_0\, \sqrt{P_{10}\, P_{20}}\, \sin\phi \quad , \qquad\qquad 6$$

where R_o is the photosensor responsivity and P_{10} and P_{20} are the optical powers corresponding to the particle displacement and force signals, respectively.

The error in position measurement is assumed to be solely from random thermal fluctuations of the particle in the trap. The photocurrent noise power, i_n^2, is therefore,

$$i_n^2 = N\, R_o^2 \left(\Delta P/\Delta x\right)_0^2 k_B T/k \qquad\qquad 7$$

where $(\Delta P/\Delta x)_o$ is the conversion factor between particle displacement and detected optical power. The photocurrent signal-to-noise ratio is next calculated,

$$SNR = \frac{i_s}{i_n} = \frac{\sqrt{N}\,\sqrt{P_{10}\,P_{20}}\,\sin\phi}{\left(\Delta P/\Delta x\right)_0 \sqrt{k_B T/k}} \qquad , \qquad\qquad 8$$

and when set equal to one, an expression for the minimum detectable phase angle ϕ_{min} can be determined. Assuming $\sin \phi \sim \phi$, the minimum detectable phase angle is

$$\phi_{min} = \frac{\left(\Delta P/\Delta x\right)_0 k_B T/k}{\sqrt{N}\sqrt{P_{10} P_{20}}} \qquad . \qquad 9$$

Setting $N = 1$, $(\Delta P/\Delta x)_o = 1$ and inserting nominal values for the other parameters (namely, $k = 5 \times 10^{-6}$ N/m, $P_{20} = 10^{-3}$ W and $P_{10} = 10^{-3}$ P_{20}) gives $\phi_{min} \sim 32$ μrad; a reasonable value. For a particle of mass 6×10^{-15} kg (a one micrometer radius glass sphere) and a modulation frequency of 30 Hz, the relative change in particle mass can be measured to within 3×10^{-4} in 1/30 s. Integrating for ten seconds increases the sensitivity to ~ 30 ppm and integrating for longer times (over many cycles) further increases the sensivity.
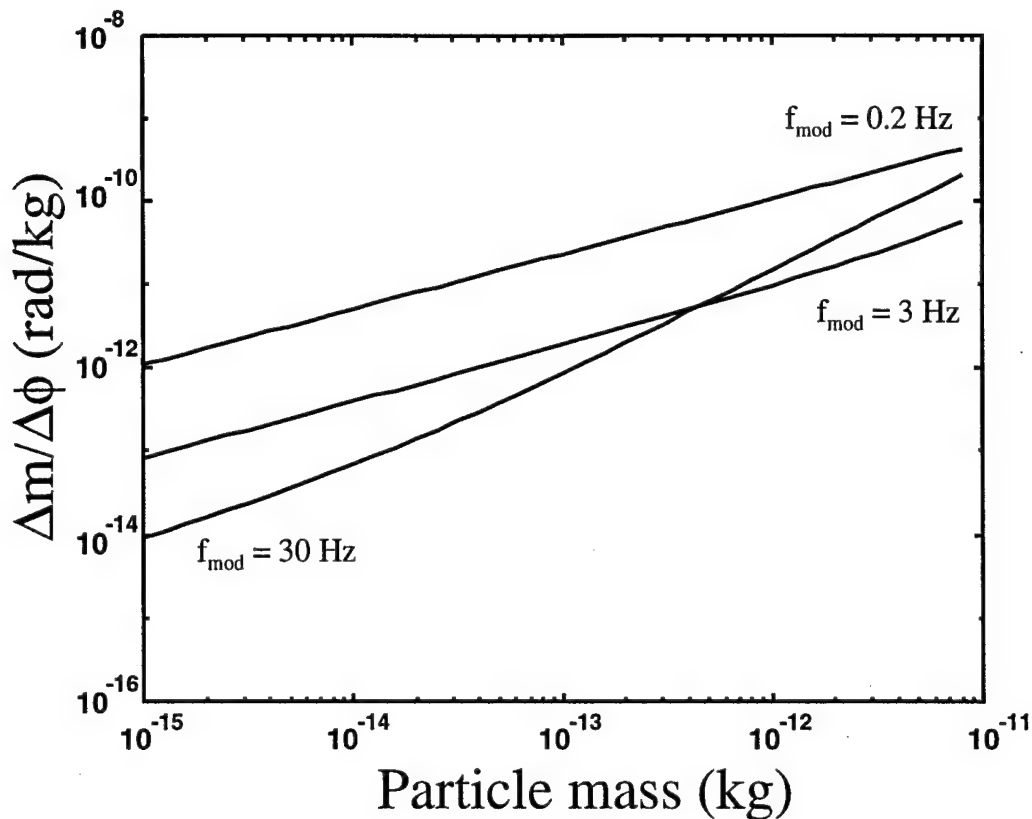
Figure 5: Phase sensitivity to particle mass change for different modulation frequencies.

3.0 Preliminary Results

As a preliminary demonstration of the technique, a photon trap apparatus was constructed according to the diagram in Figure 6. Two co-linear laser beams of different wavelength are focused onto a glass microsphere through a high numerical aperture lens. The trapping laser beam at a wavelength not absorbed by the microparticle is approximately ten times more powerful than the second, probe laser beam. Laser light reflected from the particle is passed through the beamsplitter cube and a notch filter that selectively attenuates the trapping laser beam and transmits the probe laser beam which is focused onto a photodetector. By virtue of the geometry of the optical paths traversed by the probe laser beam, axial displacement of the particle result in a change in optical intensity incident on the photosensor. Calibration of this optical arrangement correlates the particle displacement with photocurrent from the sensor. This sensor arrangement is preferentially sensitive to axial displacements since the radial trapping force is

10

substantially stronger thus limiting particle displacements transverse to the optical beam. Amplitude modulation of the trapping beam applies a perturbative axial force to the particle. The trap stiffness and force applied to the particle was calibrated with a variety of techniques which yielded similar trap stiffnesses (~5 x 10^{-6} N/m) for an incident laser power of ~ 20 mW. The corresponding axial force acting on the particle at this laser power was determined to be 2.4 pN.
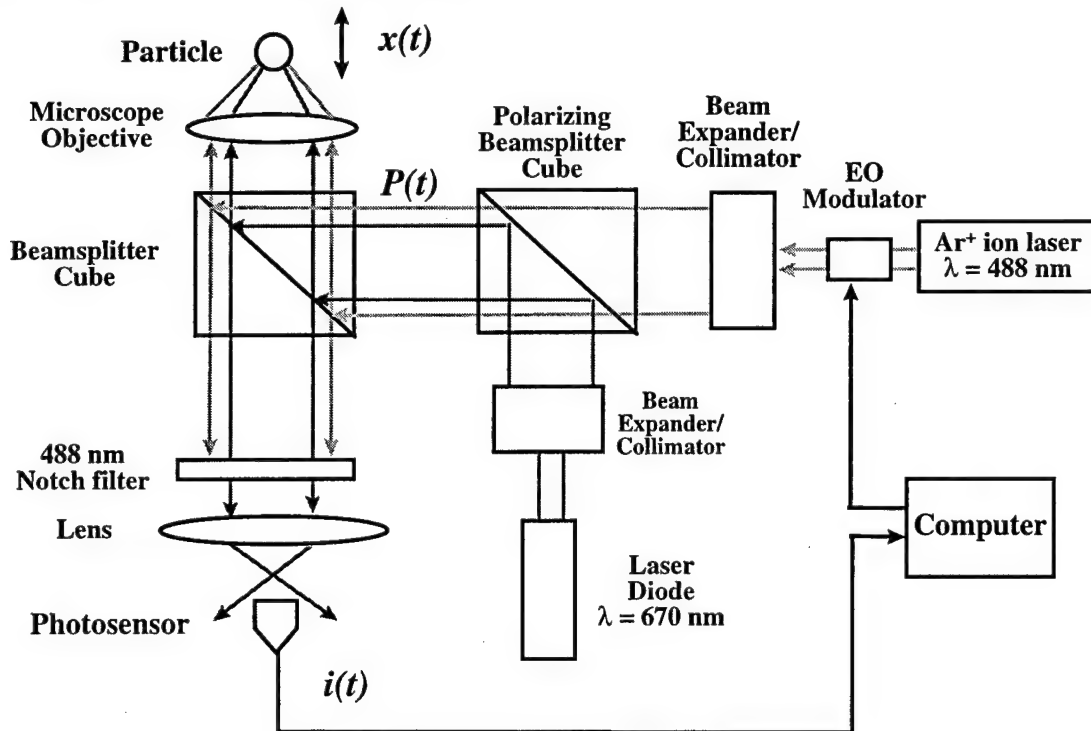

Figure 6: Experimental setup for preliminary chemical sensor measurements.

The phase transfer functions for two glass microspheres of different diameters immersed in water were measured and the results shown in Figure 7. Note the excellent agreement between theoretical prediction and experimental measurement. Furthermore, the phase transfer function clearly reveals a mass difference of only 53% between the two microparticles.
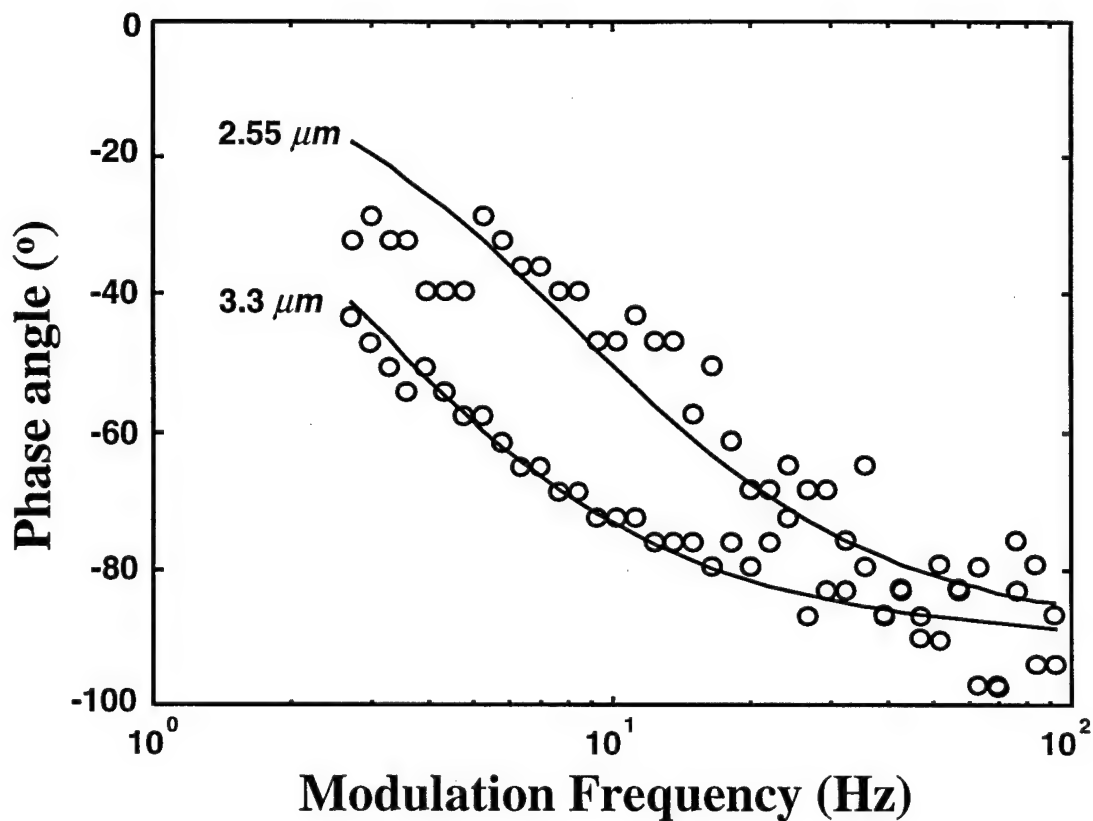
Figure 7: Experimental and theoretical phase transfer function for two glass microspheres of different diameters.

4.0 Future Directions

There are several future directions for this work. First, further experimental work is required to establish the detection sensitivity and reliability of the method using microspheres functionalized to detect specific chemical entities and be insensitive to others. This work will proceed with the experimental apparatus described in Figure 6 and investigate both liquid and air-based detection schemes.

Photon force-based Chemical Sensor

$$\frac{\Delta \omega_o'}{\omega_o'} = -\frac{1}{2}\frac{\Delta m}{m}$$

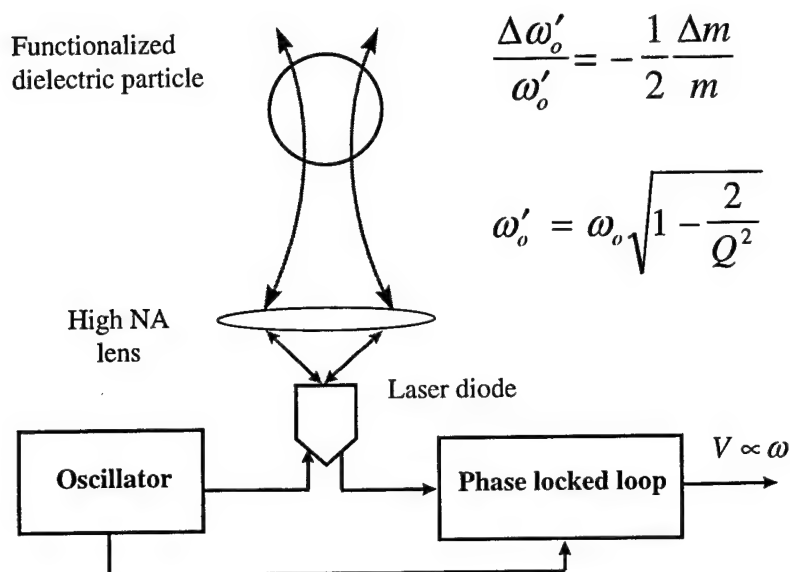$$\omega_o' = \omega_o \sqrt{1 - \frac{2}{Q^2}}$$

Figure 8: Proposed inexpensve photon force based chemical sensor.

Second, an inexpensive PF sensor design is proposed in Figure 8 based on an inexpensive, mass-produced laser diode. Light from the laser diode is focused to trap a functionalized dielectric sphere in stable force equilibrium. Laser light backscattered from the sphere is detected by the photosensor that is integrated into the laser chip and the particle position is measured based on a calibration of the photosensor output. The optical power is amplitude-modulated, the particle position recorded with time and the particle mass estimated from the relative phase difference between the modulating force function and particle displacement. Indeed, the laser cavity could be used itself as a sensitive interferometric displacement sensor to record small changes in particle position. If the particle is suspended in air rather than a liquid, Brownian motion is minimal and the sensitivity to mass changes correspondingly much higher. The smaller viscous damping forces on the particle in air as opposed to a fluid would increase the Q factor of the mechanical resonator thereby further enhancing its sensitivity. We will construct such a chemical sensor and evaluate its performance.

Physical Aids

CHAPTER 12

Personal Aid for Mobility and Monitoring: A Helping Hand for the Elderly
          S. Dubowsky

Home Automation and Health Care Consortium Report

Personal Aid for Mobility and Monitoring: A Helping Hand for the Elderly

Professor Steven Dubowsky
Department of Mechanical Engineering

Reporting period: 10/1/97 to 10/1/98

1. Motivation and Program Objective

The objective of this research program is to develop the fundamental technology for a Personal Aid for Mobility and Monitoring (PAMM) that meets the needs of elderly living independently or in senior assisted-living facilities. As an elderly individual moves toward higher levels of care (i.e., from independent living to assisted living facilities to nursing homes), costs increase and quality of life decreases rapidly. The largest change occurs during the transition into a nursing home. Delaying the onset of this transition (with PAMM's) will be extremely beneficial for the individuals and economically favorable for society.
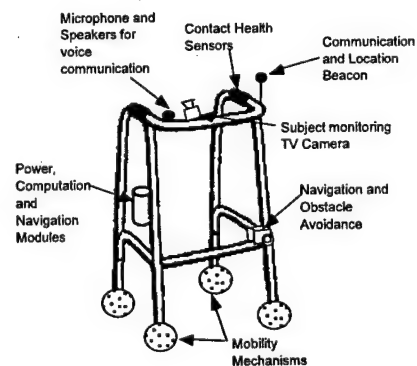
a. **Traditional**   b. **PAMM**

Figure 1: A Helping Hand for the Elderly.

2. Year I Tasks

During this first year of the program, the tasks addressed were:

- Establishing the needs of potential system users
- Defining the system's performance requirements
- Proposal of system-level concepts
- Identification of essential technical areas requiring research and development and the associated risks.

3. Users Needs

Based on studies of physical ailments in the elderly, the basic functionality of PAMM was established. Table 1 outlines the needs of the elderly and the associated physical infirmities.

Table 1. PAMM Functions

Need	Physical Deficiency	Cause
Guidance	Failing memory, disorientation	Senile dementia, including Alzheimer's.
Physical support	Muscular- skeletal frailty, instability	Osteoporosis, Diabetes, Parkinson's, Arthritis, lack of exercise, failure of vestibular organs.
Health Monitoring	Poor cardiovascular function, susceptibility to strokes and heart attacks.	Poor diet, old age, lack of exercise, illness (e.g., flu or pneumonia)
Medicine and Other Scheduling	Need to take a variety of medicines coupled with failing memory and disorientation,	Senile dementia, general failure health.

4. System Performance Requirements

A preliminary set of technical performance goals for the PAMM were defined based on the singular assumption that the system would be used inside an assisted living facility. The performance goals are summarized in Table 2.

Table 2: PAMM Performance Goals

User and operation environment characteristics	
Potential Users	Elderly with mobility difficulty due to physical frailty and/or disorientation due to aging and sickness.
Environment	Assisted-living facilities. Known structured indoor environment with random obstacles such as furniture and people. Flat and semi-hard floor or ramps less than 5 degrees.
System function and features	
Physical stability	Provide equal or better stability than that of a standard walker.
Guidance and obstacle avoidance	Provide guidance to destinations via global sensing, planning and obstacle avoidance strategies.
Health monitoring	Provide continuous health monitoring (details TBD).
Communication	Provide communication with patients and caretakers.
Mechanical specifications	
Mobility device	Compact and robust wheel-based mobility platform with design reconfigurability.
Speed	Able to assist the elderly walking up to 0.5m/s.
Loading capacity	Able to support the average body weight of an elderly person and provide 2 to 4 kg pulling force for stability and guidance.
Weight	Approximately 15 kg.
Physical size	Approximately equal to a conventional walker
Battery life	About 5 hours between charges.
Sensing and computing	
Computing power	On-board computers sufficient for planning, control, health monitoring and communication.
Sensors and aides to Navigation.	Vision based global sensing for high level planning. Ultrasonic based sensors for obstacle avoidance. Optical encoders for dead reckoning and motion control. Map based localization.

5. The PAMM System Concept

The basic PAMM concept is shown in Figure 1. It consists of a mobility aid that is able to locate itself in a facility by visually reading simple sign posts strategically placed on the ceiling of the eldercare facility. The PAMM will use acoustic sensors to provide local obstacle avoidance and enable it to maneuver in crowded environments. It will communicate with the facility central computer to obtain information such as the user's schedule and the facility's updated maps.

The battery powered near omni-directional drive will aid with mobility and navigation. It will also assist the user with physical stability. The on-board control and

3

planning systems will "learn" the characteristics of the user and adapt their behavior to be most comfortable and responsive to user. Sometimes to meet a user's needs, they will need to exert some control over the user's actions. The PAMM system will use a suite of sensors, including sensors to "feel" the force and torque interactions between the PAMM and the user. In addition, the PAMM will carry sensors to monitor the health and condition of the user and transmit their status to the facility computer system.
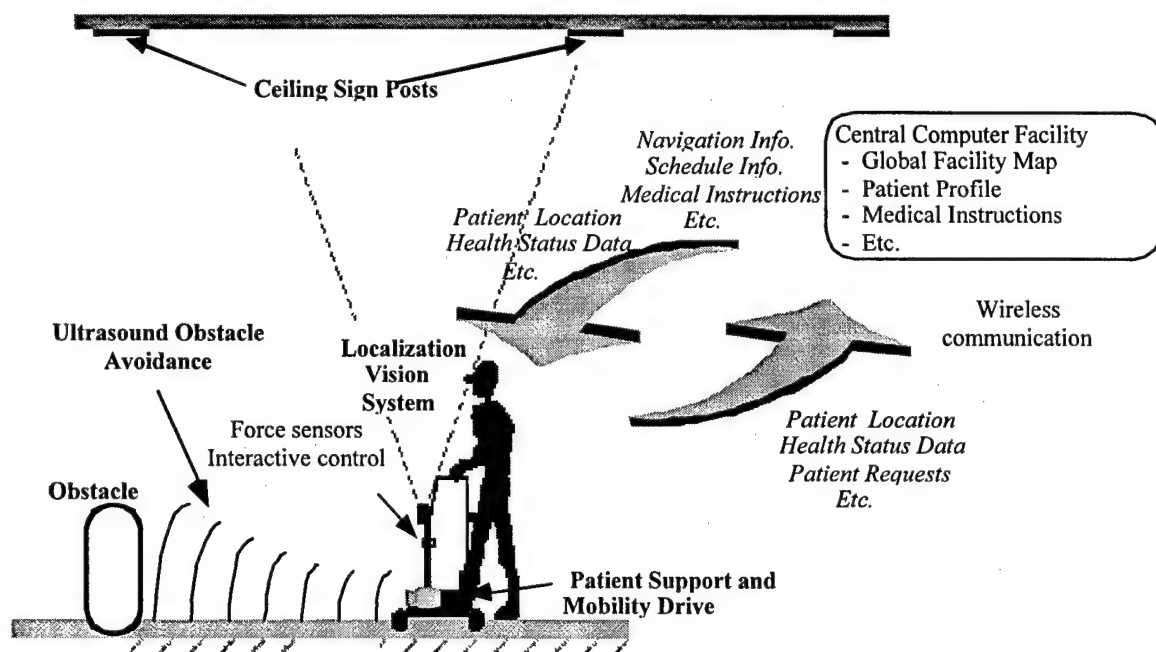

Figure 2: PAMM System Concept.

To develop these concepts a prototype system (MOD I) has been designed and constructed. It is essentially a "smart cane", and the technical details of this system are given in Appendix A. The cane is simple enough so it could be designed and fabricated quickly, yet close enough in functionality to a mobility aid so it could be used to obtain practical results. It is being used to study and evaluate many of the features and capabilities envisioned for the PAMM concept. The MOD I system,

however, cannot provide significant physical support to elderly individuals with serious mobility handicaps. During the second year, a "smart walker" (MOD II) will be developed to address these physical-support issues. However, most, if not all, of the localization, obstacle avoidance, planning and control algorithms, and health monitoring will be directly transferable to the MOD II system.

6. Technical areas requiring research and development and the associated risks.

During the past year, a number of technical challenges in the development of PAMM have been identified, and work was focused on developing conceptual solutions to these problems. A major part of the work this year was centered on designing and fabricating the PAMM MOD I system, which is described in Appendix A. This system will be used to study the effectiveness of the solutions. These challenges, solutions, and plans for studying their efficacy are summarized below. They will form the basis of the work during the coming year.

6.1 Man-Machine Interface

The man-machine interface is an area where the most challenging technical issues are expected to surface. This is because the users are expected to be prone to behavioral disorders (irrational attitude mainly due to loss of memory) or to physical problems (reduced mobility). Nearly all the work being done in this area is at very early stages, but once a prototype is operational, it will be used for testing various interfaces.

In most robot applications, the user tends to be the one who controls the amount of robot-autonomy. In the case of PAMM, however, the roles are reversed. PAMM will use all available sensory information (health, kinetics, dynamics and environment) to give or take autonomy from the user appropriately. To prevent unnecessarily abridging the independence of the user, PAMM will need an appropriate model for estimating rationality of its user's actions. Deviations from routines, for example, might be considered an indication that the user is confused. When a prototype PAMM is completed, it will be used to develop these models of rationality.

A force/torque sensor will be an essential component of PAMM's interface. Due to the complicated nature of the man-machine interaction, however, the information generated by the force/torque sensor must be analyzed before a control system that uses it can be designed. Once a prototype PAMM (i.e., the MOD I system described in Appendix A) becomes operational, it will be used to gather data on the force/torque interaction between PAMM and a user in motion. An important function of the interface is its "adaptation" to the users. Here, adaptation means the capability of the system to evolve in such a way that the various users feel "comfortable" with PAMM's help. Since behavioral disorders and physical problems vary from one user to the other, a first step will be to include in PAMM the ability to recognize its current user. Once a user is identified, a first adaptation can be done at the mechanical system level accounting for simple data such as the user's size. This part involves the reconfigurability of the physical system.

During locomotion, many challenges have to be addressed and solved by all available sensory information. The force/torque sensor data will provide useful information concerning the adequacy of PAMM's nominal speed for the current user. This requires the definition of specific "metrics" characterizing "comfort." For example, the average percentage of weight that the user puts on the cane or walker during locomotion could correspond to the degree of support assistance required. Feedback to the mechanical system to raise or reduce the height of the support points should be done accordingly. A measure of the frequency of the force/torque sensor signals over several steps is another characteristic of comfort, since standard gait pattern are cyclic.

One very challenging issue is giving PAMM the ability to accommodate for its user's behavioral disorders. Simple models of human behavior can be elaborated (using so-called "Hidden Markov Chains") to infer whether the behavior disorders are totally random or related to specific situations that the user doesn't like. Such an analysis could show, for example, that a user doesn't feel comfortable passing through a

particular door, or near a particular table. This information should be given to the planner so that it can produce trajectories that are less unsettling to the user.

6.2 Localization

In the current design, PAMM will perform dead-reckoning (DR) using motor tachometers and castor-mounted encoders. Alone, however, this scheme might not provide the desired accuracy for localization. Errors in wheel diameter, slippage, and other factors will contaminate position estimates. PAMM will therefore use a second localization technique for increasing the accuracy of the position estimates. A vision based localization system will improve PAMM's localization accuracy to the design requirements. The system will use an upward facing camera mounted on PAMM, and asymmetrical symbols on the ceiling for position and orientation information.

There are several challenges in designing this system. The algorithms must be fast enough to be done in real-time and to not interfere with other critical systems such as planning and health monitoring. To accomplish this, the algorithm will be optimized using simpler symbols, intelligent programming, and optimizing compilers. The effectiveness of this system will be measured by how accurately and quickly the vision system can localize PAMM.

In addition, it is unclear how often PAMM will need to perform it localization. This will depend, for example, on DR's sensitivity to inaccuracies in perceived motion. A simulation is being written to estimate this sensitivity and determine a minimum update frequency for localization.

6.3 Obstacle Detection and Avoidance

Since PAMM is to function in a dynamic environment, it requires an obstacle detection system in addition to basic maps of the facility. PAMM also has to navigate through doors and occasionally between two closely spaced obstacles. PAMM therefore requires the ability to track objects that are within 10 feet of it and inside an arc of about 150 °across its front end. An acoustical (ultrasonic) system is being designed to provide

this capability. While ultrasonic technology is already in use in many mobile robots, there still exist several challenges for implementing this obstacle avoidance system in PAMM. The system must be small enough so it is unobtrusive. This will involve designing a modular system with careful selection of components geared toward reuse.

Another challenge is integrating the obstacle detection with the man-machine interface. The actions PAMM takes in response to an obstacle, for example, will depend on the whether PAMM is providing guidance or just physical support to the user. This is one of the issues that will be addressed as PAMM's man-machine interface is developed.

The obstacle avoidance system must also be integrated with the planning system. The detected obstacles will be compared with the facility maps to generate current maps. In some cases, obstacles will already be in the maps, but in other cases, the maps will have to be updated. These current maps will assist in the high-level planning of PAMM. It is also desirable to have some sort of "reflexive" control where the obstacle avoidance system interacts directly with the control system when the safety of the user is in jeopardy.

6.4 Low Level Control

One of the major challenges in developing PAMM lies in designing an admittance-based controller that works despite the non-holonomic nature of the device. The singularities caused by the system's non-holonomics make it difficult to map a desired "stiffness" at the handle to a "stiffness" in the actuators. The actual mobility characteristics of PAMM will be studied, and different motion control methodologies including the non-holonomic feedback control will be implemented and tested.

For trajectory control, PAMM will rely on dead reckoning corroborated by a localization system. For improved dead reckoning accuracy, PAMM will use strategically placed passive-castor encoders along with the motor-encoders. To accurately integrate the encoder information into the dead-reckoning system the kinematics of the passive-castors needs to be investigated.

6.5 Design Issues

Another major conceptual challenge will be the development of the PAMM module design, from essentially a guidance aid to a system that offers substantial physical support, such as a walker. Active reconfiguration of the module for stability, tip-over avoidance, and maneuverability is highly desired. Contact friction becomes an important criterion, as the walker must not slip under any horizontal loads. Finally, overall weight and ease of storage are product design issues that concern not just the user, but those who take care of the user.

Other system components will be retained and improved upon from MOD I. Additional features that are considered for MOD II include:

- Learning controls programs that remember user behavior and adjusts man-machine interactions accordingly.
- Learning localization program so that the walker can configure itself to a new environment.
- A self-adjusting camera lens with a range of lens angles to accommodate different ceiling heights.
- Active communication between PAMM and its user, enabling a "parked" walker, for example, to return to its owner upon command.
- Other features that aid the elderly, disabled, and their caretakers.

Appendix A - MOD I System Design

A. 1 Physical Design

Figure A.1 shows the design of the smart-cane. It consists of a mobility platform, an array of ultrasonic sensors for obstacle detection, a suite of health sensors, a six-axis force/torque sensor, a vision system for localization (not shown), and an on-board computer with interface electronics. The main considerations in the selection of an appropriate drive and steering configuration for the mobility platform were maneuverability, controllability, traction and stability, navigation, environment impact and simplicity.
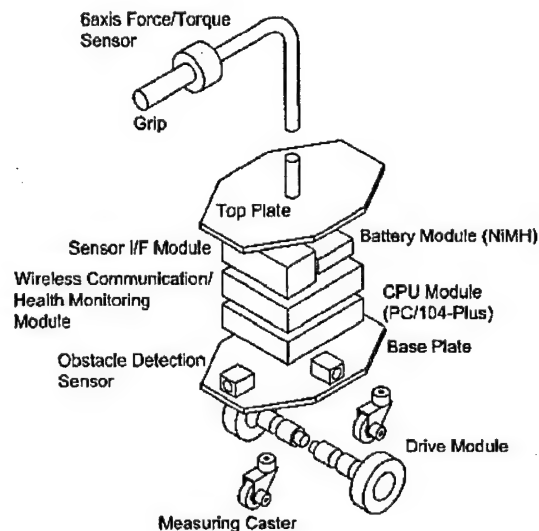
Figure A.1:   Smart-cane Design.

Various mobility forms and wheel configurations were investigated. The MOD I system uses two-wheel skid-steering drive because of its simplicity in construction. It has two individually controlled driving wheels and up to two passive casters, as shown Figure A.2. This configuration also has relatively good maneuverability in congested environments as it allows an on-spot spin. Each drive motor has an incremental optical encoder for motion control and dead reckoning.

There will be some slippage of the driving wheels. This, along with other sources of error, limit accuracy of the dead reckoning based solely on the encoders on the driving wheels. Additional encoders on the passive castor wheels and their vertical shafts measure the castor rotation and heading. This will be used to improve the navigation accuracy of the system. The mobility base is modular, so the castor or motor assemblies can be rearranged to study different configurations. The front castor, for example, could be removed and the motor assemblies could be moved forward into a triangular configuration that might improve wheel traction on uneven surfaces.
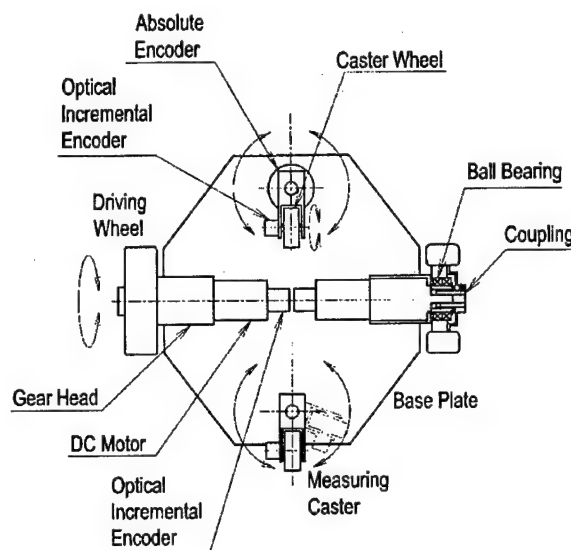

Figure A.2: Layout of Motors, Casters and Sensors.

A.2 Vision Based Localization Approach

To provide periodic updates of the PAMM location and heading, a vision-based localization, illustrated in Figure A.3, is incorporated into the design. The system will monitor signposts located at known locations on the ceilings of the facility. The signposts are placed so that at least one will always be in view of the vision system. This prevents PAMM from ever being "lost" in the environment.

The signpost, an example of which is shown in see Figure A.4, encodes its location and local "compass directions." From a captured image of the signpost, PAMM determines its position and heading relative to the signpost. The centroid of the large circular

centerpiece marker in the image provides the relative distance of the PAMM from the signpost. The position of the centroid of the smaller orientation marker relative to the centerpiece marker gives the local heading information. Finally, a group of small identification markers surrounding the centerpiece label the marker with a unique number using binary encoding. Five identification markers, for example, will suffice for a 31-signpost system. The system can be expanded to 127 unique signposts by using a seven-marker identification system.
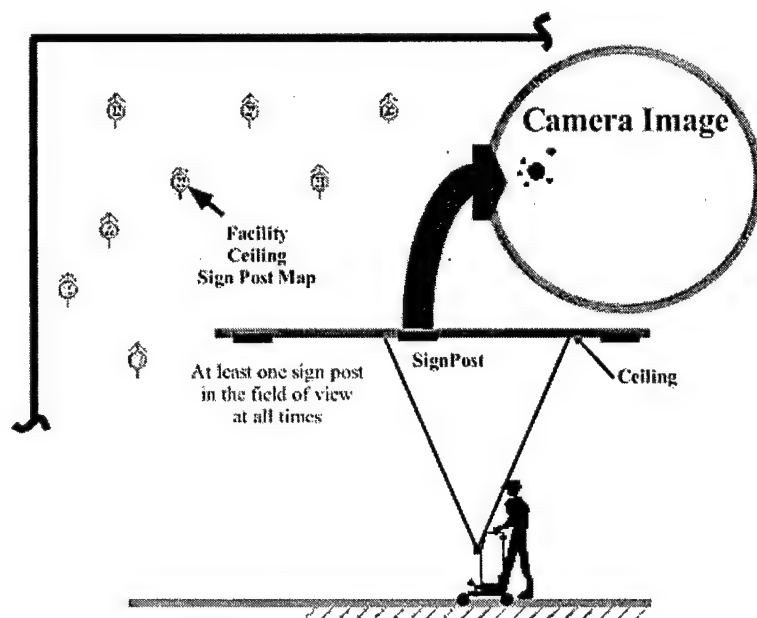

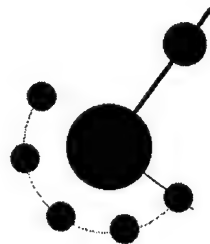**Figure A. 3: Vision Localization Approach.**

Figure A. 4 Localization Sign Post

A.3   Planning and Control

The MOD I planner will have several modes. In the passive mode, the planner acts as an observer, watching the health and condition of the user. It will only interfere if a problem is encountered. In the first active mode, the planner also acts as a guide. It might generate a schedule of destinations for the user and according to this schedule plan a sequence of paths. Due to the non-holonomic nature of the system, some non-holonomic path planing may be required.

Figure A.5 shows an overview of PAMM's planning and control system with the latter being enclosed by the dotted box. The control system ensures the smart-cane follows the trajectories from the planner. It also responds to user inputs and environment changes to provide smooth and safe motion. The controller therefore consists of two important functional parts, the admittance control model and the motion controller. The novel technique of PAMM's control system design is its admittance control.
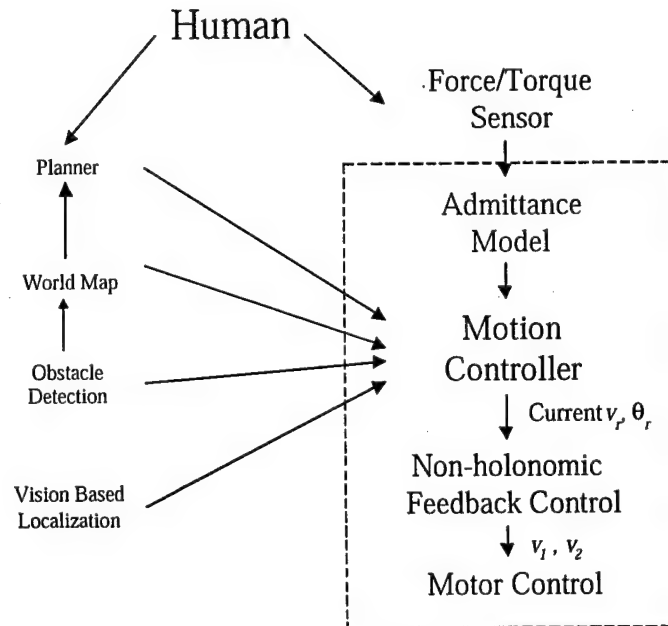

Figure A.5.   The control system overview

The admittance-based control uses the force/torque sensor mounted on the handle to measure the user/system interactions, and adjusts the control to provide a natural-

feeling interaction. In theory, the admittance controller can adjust its parameters to adapt to the user's character.

Different motion planning and control methodologies, including the non-holonomic feedback control and planners will be implemented and tested as needed.

A.4   Man-Machine Interface

Conceptual designs of the user-machine interface based on the force/torque sensor signals have been developed. The objective is to obtain a deeper understanding of the mechanics of stable guidance of the elderly. Ideally, the PAMM should be able to judge the user's mental clarity and intention and lead the user to the destination. This high level of capability will not be incorporated in the MOD I system. For the MOD I system, three modes of user/machine interaction are used. In the first mode, the user will decide the path and the PAMM will passively aid the user by giving physical support. The controller adjusts the speed of the PAMM based on the users walking patterns measured by the force/torque sensor. In the second mode, the PAMM will lead the user through a predefined obstacle-free path and give the user physical support. In the last mode, PAMM will work with the obstacle avoidance system to lead the user to a destination through an environment with unmapped obstacles.

A.5   Electronics Design

Figure A.6 illustrates the electronics architecture of the MOD I system. Two NiCd batteries, each providing 1.4Ah at 9.6V, are used as the power source. Several DC/DC converter units regulate supply voltages. The system uses two 30W DC brush motors to drive the wheels. A 1024 pules per revolution optical encoder is mounted on each motor to detect the motion. A 16-bit A/D converter acquires output signals from the force/torque sensor.
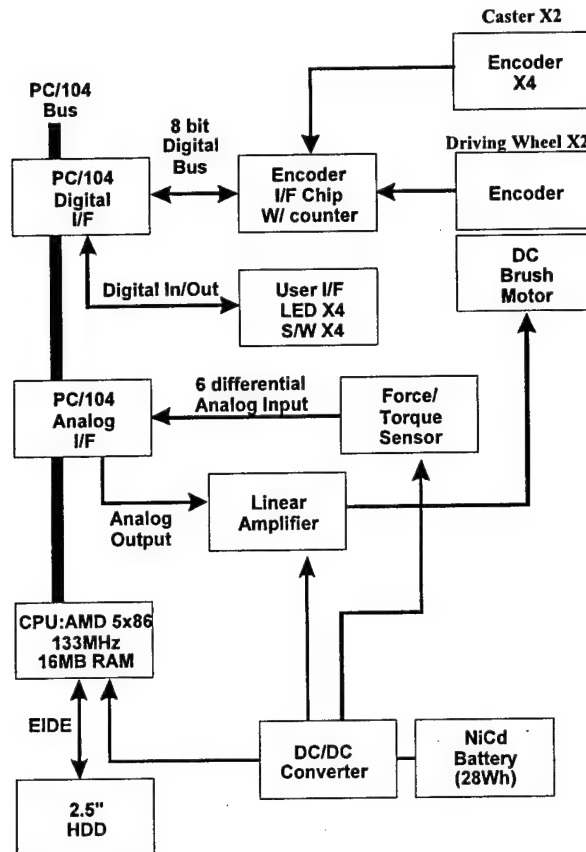
14

Figure A.6. MOD I Electronics Configuration.

A.6 Computer Architecture Overview

A PC/104-plus based computer has been chosen for MOD I system. It is a relatively cheap, powerful, and physically robust family of components. Each board simply screws on top of the next, and the stack-through header connectors provide good electrical contact. A number of modules are available that are compatible with the PC/104: CPU modules from the 8088 to Pentium II, DSP coprocessors, analog and digital I/O cards, and video-frame grabbers. PC/104-plus also supports a high-speed 32-bit bus system, which is electrically equivalent to the widely accepted PCI bus.

The PC/104 CPU module that is used has a 133MHz AMD 5x86 processor and 16MB EDO-RAM. It is also equipped with a VGA controller for use with a monitor. The CPU module can boot standard operating systems (i.e., Windows 98, NT 4.0 or DOS). Since it requires only a 5V-supply voltage, it uses a compact and simple power supply. It

also supports a low-power feature, requiring only 8W under typical operation. This makes it possible to run the system from a battery for long periods. The computer subsystem fits into a rectangular volume of 8 x 5.5 x 2 inches.

PAMM's PC/104 stacks three PC/104 modules. The first card is an analog I/F card, which provides 8 differential analog inputs (16-bit) and four analog outputs (12-bit). The second card is a digital I/F, which supports 48 digital I/O lines. The third module is a frame-grabber for the vision system. The MOD I software runs as a DOS application and is built using an off-the-shelf 16-bit C++ compiler (Borland C++).

Physical Aids

CHAPTER 13

Design and Control of an Active Mattress for Moving Bedridden Patients
H. Asada, W. Finger

d'Arbeloff Laboratory for Information Systems and Technology
MIT

Design and Control of an Active Mattress
for Moving Bedridden Patients

William H. Finger
H. Harry Asada

September 30th, 1998
Consortium Progress Report

d'Arbeloff Lab
Mechanical Engineering
MIT
Cambridge, MA 02139

Abstract

A new mechanism for transporting a bedridden patient in an arbitrary direction while lying comfortably on the bed is developed. A wave-like periodic motion is generated on a mattress surface by activating the individual coil springs comprising the mattress. The whole or part of the patient body is moved by this periodic surface movement. Varying the periodic trajectory and coordination pattern yields various movements of the patient, i.e. translation and rotation of the whole body, changing the posture of the limbs, etc. First, functional requirements for active mattresses are provided, and a prototype system is designed and built. A variety of control algorithms are developed for moving a patient in various ways. Periodic trajectory and coordination patterns are optimized in order to move the patient smoothly despite uncertainties in load distribution and actuator dynamics. Experiments using a prototype mattress demonstrate smooth body motion in both the x and y directions and rotation within the plane of the mattress surface.

1 Introduction

Patient mobility is a growing concern for health care facilities in the United States. In nursing homes alone, there are 1.5 million bedridden patients [1]. These patients require assistance for even simple tasks, and must be turned every two hours to prevent decubitous ulcers, or bedsores, from forming. 25% of all patients in nursing homes will suffer from these painful sores [2].

In order to meet the needs of the patients, caregivers are forced to move the patients manually, a very labor intensive task. In terms of Lost Work Day injuries, this task made being a nursing home employee

1

the most dangerous service job in America in 1996, with 8.2 full time employees losing a day for every 100, on average [2]. Many of these injuries are to the back or shoulder; 1 in 22 nursing assistants lost a workday from a back or shoulder injury in 1994 [3].

In the past, simple mechanical devices and equipment have been used to transfer patients between beds and wheelchairs, for example hoists or belt conveyers. To eliminate this bed-chair transfer, a hybrid bed/chair system, RHOMBUS, has been developed and tested [6]. Recently an active bed with a mechanical linkage mechanism succeeded in moving a human lying on the bed by creating a surface wave in one direction [5]. Despite these endeavors, aids for the bedridden that provide diverse functionality have not been developed. Patients must be moved two dimensionally, and individual limbs must be moved, while the patient lies comfortably upon the bed.

The goal of this paper is to develop a new type of active bed that provides both high mobility and sleep comfort. Special actuators imbedded into a mattress generate periodic surface movements that transfer the bedridden in an arbitrary direction.

A prototype has been constructed that illustrates how these goals might be met. A comfortable bed surface was obtained by making use of the springs available in a commercial mattress. These springs form the nodes that are the basic unit of control on the bed surface. Smooth motion was obtained by making careful selection of the trajectory taken by these nodes during the motion.
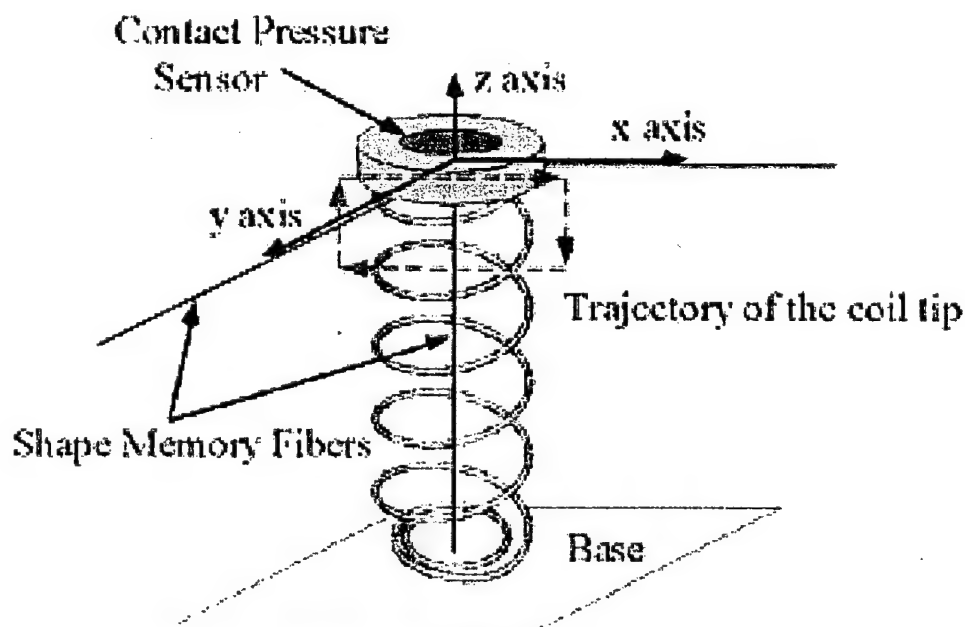

Figure 1: Actuated Coil Spring

2 Design

2.1 Functional Requirements

The objective of the active mattress is to assist caregivers in positioning and transporting bedridden patients. The mattress must be comfortable and appropriate for long-term care. Accomplishing this objective requires that the following functional requirements be met:

1) The bed surface must provide at least the same level of comfort as that of home-use beds,

2) The whole patient body must be moved in an arbitrary direction on the bed surface without lifting the body,

3) The limbs of the patient must be moved individually while lying on the bed, and

4) The motion must be smooth, with minimal jerk and disturbance of the patient.

To meet the first requirement, we employed coil springs used in a standard commercial mattress to support the patient. To meet the second and third requirements, these coil springs are instrumented and activated, as shown in Figure 1. Namely, the tip of each coil spring is moved horizontally in both the x and y directions as well as vertically in the z direction, through the use of actuators imbedded in the mattress. The coordination of the motion of the individual coil springs creates a variety of body movements, including both whole body transfer and individual limb movements. Furthermore, these active springs are controlled in such a way that the jerk and disturbance induced by the activation of the coil springs may be minimized while moving the patient body, satisfying the last functional requirement. When these coil springs are not activated, the mattress becomes a standard, passive mattress whose coil springs were tuned to maximum sleep comfort by the bed manufacturer. Therefore the same level of comfort as the original mattress can be retained.

2.2 Prototype

Figure 2 shows the first prototype of the active mattress to meet the functional requirements described above. The mattress contains 32 active coil springs taken from a commercial bed. Each spring moves independently in the z direction with a shape memory alloy fiber actuator connecting the top plate of the spring to the base frame of the mattress. Shape memory alloy fibers are compact and powerful, as well as inexpensive compared with traditional actuators. Their limitations are low bandwidth and nonlinear behavior. Since in the prototype the springs need only detach and reattach with the body, and on-off motion, they suffice for the initial design.
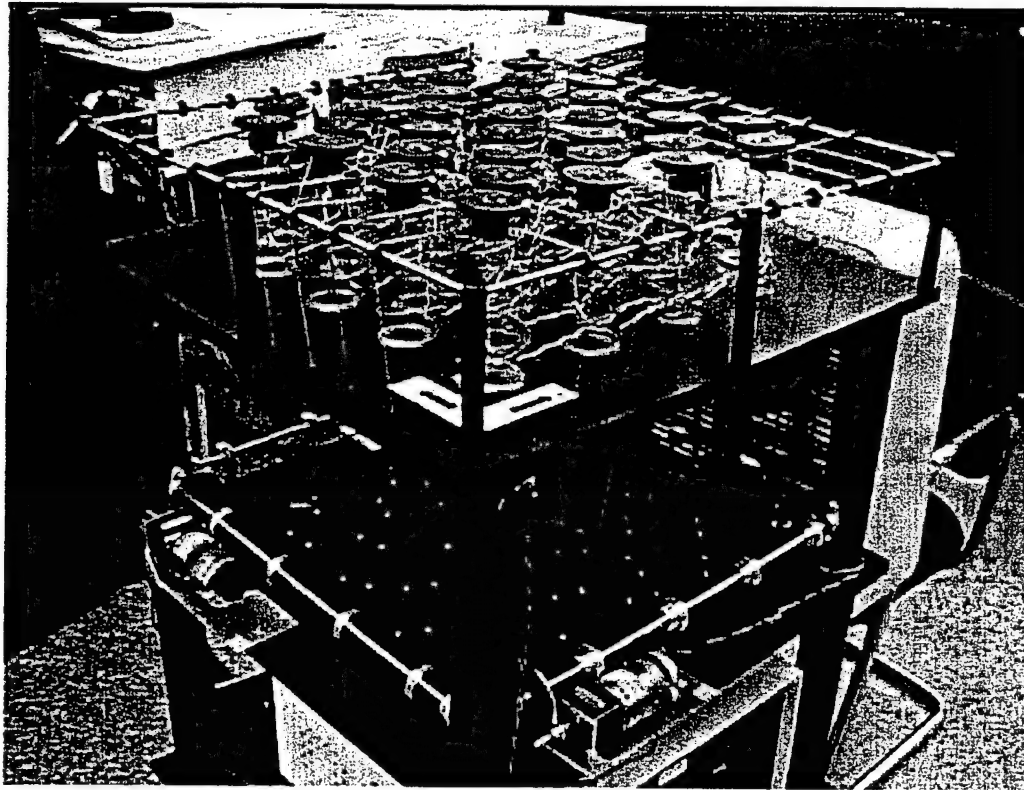
Figure 2: Surface Wave Prototype

Conversely, the horizontal motion needs higher accuracy and faster speed of response than that of the z axis. Since the coil springs need not be moved independently, the set of springs will be divided into several subsets and actuated together. The functional requirements described in section 2.1 may be met by a few independent actuators driving all springs in the x and y directions. In the prototype mattress shown in Figure 2, two independent servo motors are used for driving each horizontal axis. As shown in Figure 3, every second coil spring along the x axis is connected by cables and moved together by the same actuator, yielding displacement Δx_1, while the remaining nodes are moved by a second actuator an amount Δx_2. Likewise, two actuators in the y direction provide displacements Δy_1 and Δy_2, respectively.

The coordination of these four actuators with the z axis actuators creates a variety of periodic motions on the mattress surface that allow the human body to move in an arbitrary direction. For example, to move the body in the +x direction, we move one set of springs in the +x direction, detach it from the body, move it in the –x direction to its starting position, and then reattach. If the second set of springs moves 180° out of phase with this motion, the body may be moved continually in the x direction. Combined with similar motions in the y direction we can generate body motion in arbitrary directions. Furthermore, changing the state of z-axis actuators in coordination with the horizontal axes can cause the body to rotate, or generates localized body motions, as will be addressed in the following sections.
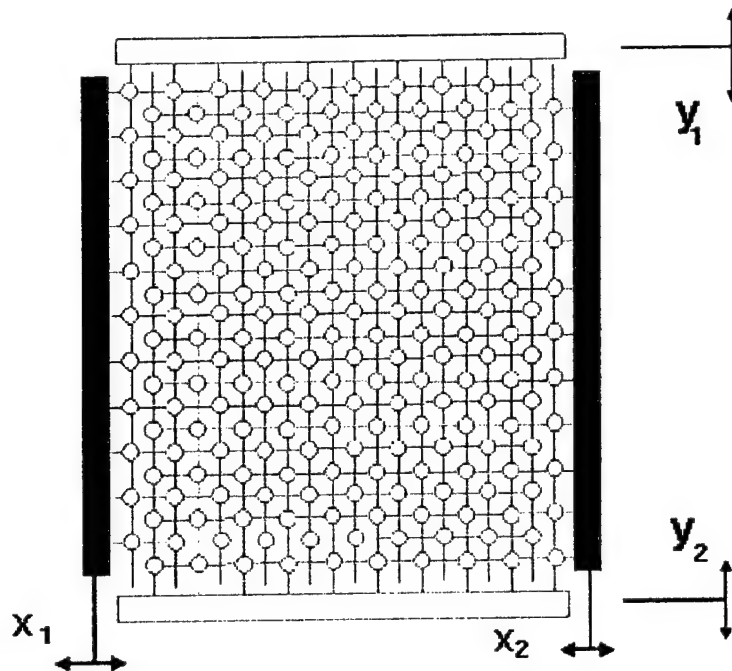
Figure 3: Alternating Grid of Nodes

3 General Control Algorithms

3.1 Notation

The control algorithm briefly described in the previous section is generalized and formally presented in this section. It will be shown that diverse motions of the human body can be created by coordinating horizontal and vertical motions in different modes. To represent diverse motions in a unified manner, we introduce the following notation.
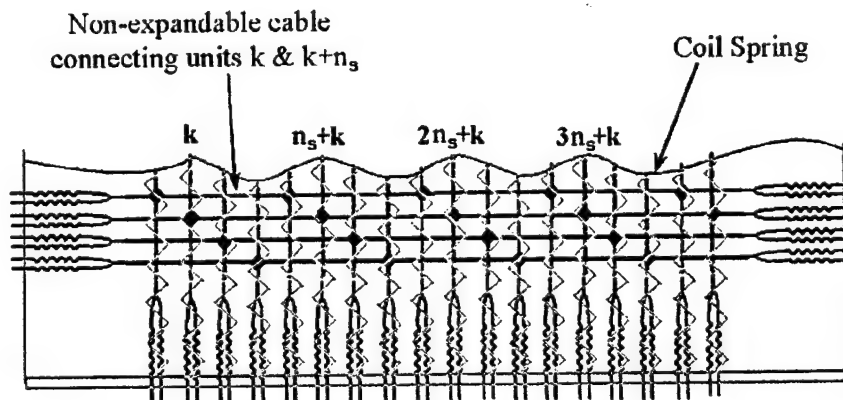

Figure 4: Node Connections

Figure 4 shows the schematic of the active mattress system. For the sake of simplicity, the figure shows a single array of active springs along the x axis, although the actual mattress is two-dimensional. Each

active spring, termed a "node", is numbered 1 through n. The coordinates of the tip of the kth node are denoted (x^k, y^k, z^k), $k = 1, \ldots, n$.

As mentioned previously, each node is controlled individually in the z direction to take either a high or low position:

$$z^k = \begin{cases} 1: \text{high position attached to the body} \\ 0: \text{low, detached from the body} \end{cases} \tag{1}$$

The horizontal coordinates of the kth node are varied from its unforced position, $\bar{x}^k$ and $\bar{y}^k$, to a deflected position by the horizontal actuators driving the kth node.

$$\begin{aligned} x^k &= \bar{x}^k + \Delta x_i \\ y^k &= \bar{y}^k + \Delta y_i \end{aligned} \tag{2}$$

Where Δx_i and Δy_i are deflections generated by the ith actuators of the x and y axes, respectively. As mentioned before, nodes are grouped together for horizontal movements, and are driven by several independent actuators. Let N be the number of groups, or *node sets*, and S_i be the ith node set, $1 \le i \le N$, containing all the node numbers, k, of the nodes moved simultaneously by the same horizontal actuators.

$$S_i^x = \{ k \mid \text{All nodes connected to the } i^{th} \text{ horizontal actuator} \} \tag{3}$$

Note that two actuators are used for moving the nodes in S_i in both the x and y axes. Note also that any actuator node sets S_i and S_j are exclusive, and every node belongs to one and only one actuator node set.

$$S_i \cap S_j = \phi \quad \forall i \ne j \tag{4}$$

$$S_1 \cup S_2 \cup \ldots \cup S_N = S \tag{5}$$

Where $S = \{1, 2, \ldots, n\}$ and ϕ is the empty set. In general, the number of nodes per node set n_S is a function of the number of actuators per axis N and the number of nodes n:

$$n_S = \frac{n}{N} \tag{6}$$

The functionality obtained as a function of the number of actuators per horizontal axis is given in Table 1. The algorithms used to perform these tasks are given in the next section.

Functionality	N
Intermittent Translation	2
Intermittent Rotation	2
Smooth Translation	2
Smooth Rotation	4
Intermittent Localized Translation	2
Smooth Localized Translation	3
Smooth Localized Rotation	5

Table 1: Number of Horizontal Actuators per Axis _N_ as a Function of Desired Functionality

2.2   Algorithms

Whole Body Translation

In this case we wish to move a body in the direction of angle θ from the x axis. The node sets are coordinated such that:

$$\Delta x_i = h(t_i)\cos\theta \tag{7}$$

$$\Delta y_i = h(t_i)\sin\theta \tag{8}$$

where $h(t_i)$ is a periodic continuous function of time t_i that generates a reciprocative motion in the horizontal direction. All node sets move along the same trajectory, and therefore use the same function h, but have different phase angles. The time t_i is given by:

$$t_i = t + T\frac{\phi_i}{2\pi} \tag{9}$$

Where T is the period of the function h, and ϕ_i is the phase differential between node i and i-1. The nodes will typically be equally spaced in phase:

$$\phi_i = \frac{2\pi}{N}(i-1), i = 1,...,N. \tag{10}$$

Figure 4 shows a hypothetical function $h(t_i)$. The function is monotonically increasing for the period of T_I and monotonically decreasing for the period of T_D. It shows three wave forms, with 120° of phase between them. The vertical motion of each node is synchronized with the horizontal motion so that the node is attached to the body only when the function $h(t_i)$ is increasing. In this way only forward motion is transmitted to the body.
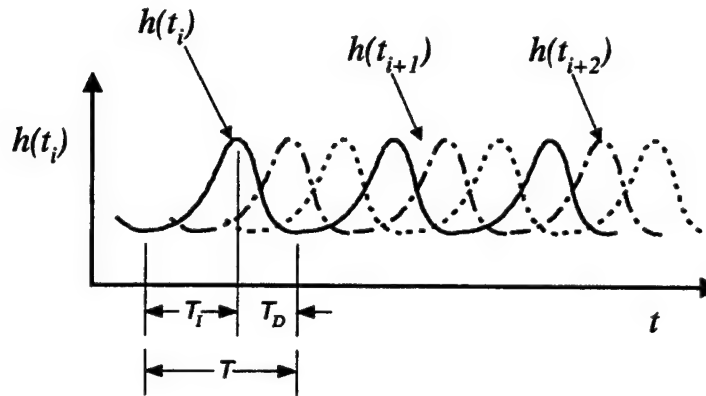
Figure 5: Hypothetical Reciprocative Function

If the kth node belongs to the ith node set, the z coordinate of this node will be represented by:

$$z^k(t_i) = \begin{cases} Z(t_i) & \forall k \in S_i \cap S^A \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

where S^A is the set of nodes located under the body A, and Z is a periodic function which depends on the trajectory used. Note that the body is always supported by at least one node set. Nodes which are not under the body are not actuated in order to save energy.

Local Movements

This algorithm allows one or more parts of the body to remain stationary, while one or more parts are moved across the bed surface. To accomplish this, at least one node set is moved in the same was as the Whole Body Translation described above; the remaining node sets are kept stationary in the horizontal direction, so that the body is supported from beneath at all times and the rest of the body is held stationary. To combine these two functional requirements in a non-conflicting manner, the z coordinate of each node is controlled in such a way that:

- The nodes that are moving horizontally and are beneath the part of the body to be kept stationary are detached from the body,

- The nodes that are moving forward and are beneath the part of the body to be moved are attached,

- The nodes that are moving backward in the horizontal direction are detached from the body, and

- The nodes kept stationary in the horizontal directions are attached to the body if the nodes are beneath the part of the body to be held stationary, or if there is no other node set available to support the part of the body to be moved.

Let us define a set S_A as all nodes under the part of the body to be moved, and S_B as all nodes under the part of the body to remain stationary. Let us assume we have $i = 1...N_S$ stationary node sets, and $j = N_S+1...N$ moving node sets. We then obtain the following expressions for this algorithm:

$$\text{For } k \in S_i, \ 1 \le i \le N_S :$$
$$\Delta x_i = \Delta y_i = 0$$
$$z^k = \begin{cases} 1 & k \in S_B \text{ or } k \in S_A \text{ and } z^j = 0 \\ & \forall j \in S_{N_S+1} \cup ... \cup S_N \\ 0 & k \in S_A \text{ and } \exists j \in S_{N_S+1} \cup ... \\ & \cup S_N \text{ such that } z^j = 1 \\ 1 & \text{otherwise} \end{cases} \tag{12}$$

$$\text{For } k \in S_i, \ N_S + 1 \le i \le N :$$
$$\Delta x_i = h(t_i)\cos\theta$$
$$\Delta y_i = h(t_i)\sin\theta$$
$$z^k = \begin{cases} Z(t_i) & k \in S^A \\ 0 & k \in S_B \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

In general, this requires one additional actuator per axis than is required to perform a given motion; this actuator remains stationary, and its nodes stay in contact with the stationary part of the body, while remaining detached from the moving part or parts of the body. In the special case of $N = 2$, the stationary node set is required to periodically support the body, while the single moving node set detaches from the body.

Moving Two Parts of the Body in Opposite Directions

By modifying the above local movements, two parts of the body can be moved in two opposite directions, such as the legs opened or closed. Assume that body segments A and B are to be moved in opposite directions to each other. The z axis motion of the nodes beneath A, $k \in S_A$, and the ones beneath B, $k \in S_B$, are set $180°$ out of phase so that forward motion alone may be transmitted to body A while only motion in the opposite direction is transmitted to B. For $i = 1...N_S$:

$$z_k = \begin{cases} Z(t_i) & \forall k \text{ s.t. } k \in S^A, k \in S_i \\ 1 - Z(t_i) & \forall k \text{ s.t. } k \in S^B, k \in S_i \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

A special case of this algorithm allows us to rotate the entire body. This is done by creating two anti-parallel velocities, symmetric to and at equal distances from the center of mass of the body, perpendicular to its longitudinal axis.

4 Nodal Trajectory Considerations

4.1 Trajectory Design

The active mattress moves a patient body through periodic and intermittent movements of an array of active nodes. To generate continuous and smooth movements, an array of nodes must be coordinated with each other, and the trajectory of each node motion must be tailored so that smooth transitions may be accomplished.

While achieving these goals, we endeavor to move the patient as quickly as possible, limited by safety and patient comfort.

One of the main difficulties in achieving these goals has been the limited performance of the shape memory alloy actuators. They are slow to respond, and their response depends on non-linear heat transfer phenomena which vary depending on ambient conditions. Therefore it is difficult to use them accurately in an open loop configuration, so that the time for reconnection and disconnection with the body is uncertain.

The node trajectory and coordination method must be optimized so that smooth patient transfer can occur despite actuator limitations. We consider the following trajectories.
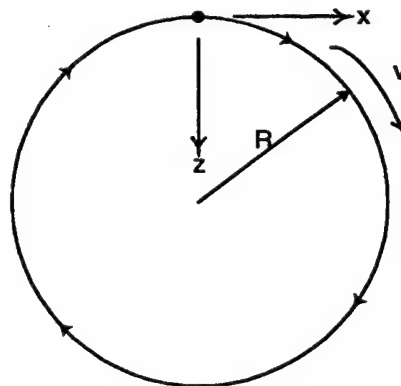
Circular

Figure 6: Circular Trajectory

Figure 6 shows a circular trajectory in a vertical plane. Without loss of generality, we assume the place in parallel to the x axis. The periodic function $h(t_i)$ introduced in section 3.2 is simply the projection of this circular trajectory onto the horizontal plane. Namely, $h(t_i) = R \sin(\omega t_i)$.

The circular trajectory is fairly smooth during horizontal motions, as long as the number of node sets N is large and the phase difference between adjacent node sets is small. However, as N becomes smaller, there is an undesirable z-axis motion created by the transfer of weight from one set of nodes to another. The total magnitude of this motion is dependent on the phase difference $\Delta\phi$ between the nodes and the radius of the trajectory, as shown by Equation 15:

$$\Delta Z = R\left(1 - \cos\left(\frac{\Delta\phi}{2}\right)\right) \tag{15}$$

where ϕ is the phase lag between adjacent nodes, and R is the radius of the trajectory. The circular trajectory can be created using a piston and crank mechanism, and was used for the original surface wave actuator [5]. However, since we make use of SMA fiber actuators, this trajectory would be highly distorted, and because the z-axis perturbation which occurs with $N = 2$ is quite significant, this trajectory is not desirable for our prototype.
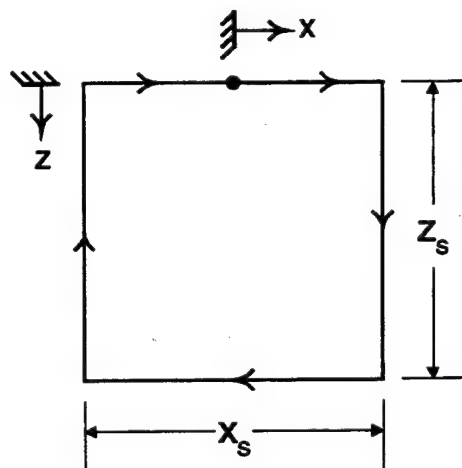
Rectangular

Figure 7: Rectangular Trajectory Geometry

To eliminate this unwanted vertical perturbation, we can employ a rectangular trajectory, as shown in Figure 7. This trajectory causes no vertical perturbation, since the path in contact with the body is a straight line. Another advantage to this trajectory is its simplicity. It can be created very easily using decoupled axes. Also, with two node sets at 180° of phase, the nodes are always travelling in opposite directions with this trajectory, so only one actuator per axis is needed. The node sets will be geared so

that they move in opposite directions. The simplicity of this trajectory allows it to perform complex maneuvers that might otherwise require many actuators per axis.

The disadvantage of this trajectory is that at the end of each horizontal motion the node must come to a stop, and wait for the adjacent node to complete its motion. Also, at the end of each vertical motion the node must be accelerated quickly, inevitably causing jerky motion. Therefore this trajectory violates our functional requirement (4).

Trapezoidal

To overcome the limitations of the previous trajectories, a hybrid of sorts has been designed. This trajectory has the benefits of zero vertical perturbation of the body, due to the straight line body contact path, combined with smooth horizontal motion. The horizontal motion is smooth since before the node makes contact with the body, it has obtained the velocity of the body; therefore, the body need not come to a stop for reconnection to occur. The same occurs during disconnection, but in reverse; the node disconnects, and then begins to change its horizontal velocity.
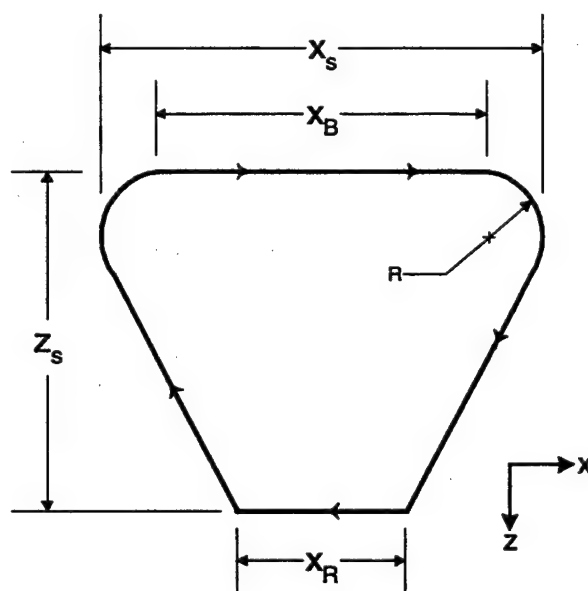

Figure 8: Trapezoidal Trajectory

Since this trajectory meets our functional requirements, we adopt it and develop a complete control method for coordinating multiple node sets.

3.2   Coordination Control

In order to achieve these connections for the trapezoidal trajectory, feedback is required as to whether the nodes are currently in contact. This is obtained using force sensors located at the interface between the

nodes and the body. Reconnection can only be estimated using time data, since feedback of vertical position is not available.
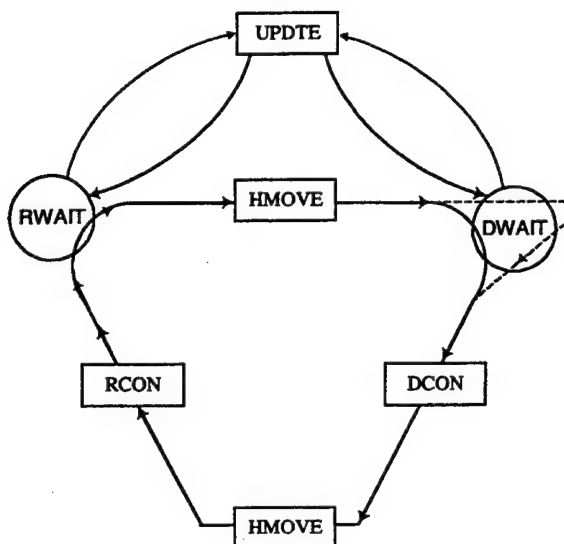

Figure 9: Trapezoidal Trajectory States

Figure 9 shows the discrete state network used, laid over the continuous path traced by the nodes. The labels in rectangular boxes represent discrete states, while circular blocks indicate conditions for state transition The DWAIT transition indicates the node is waiting for this node to disconnect, while RWAIT indicates we are waiting for it to reconnect. HMOVE and UPDTE correspond to horizontal movements and update states, respectively.

We begin with horizontal motion in contact with the body (HMOVE). The body is moved until two conditions are met: we are near the end of the horizontal stroke, and the node set adjacent to this one—and out of phase by 180° for N = 2—is in the HMOVE state.

The node set then enters the DWAIT state, where it waits for itself to disconnect from the body. The nodes continue to move at the velocity of the body. When the nodes have all disconnected, or a maximum time has elapsed, the node set enters a disconnection state (DCON). Note the dotted line on the figure. This optional path may be traced by either the HMOVE or DWAIT states, and occurs when the conditions for state transition have not yet occurred, but since we are still in contact with the body, we need to continue at constant speed.

The DCON state reverses the direction of the node set motion. When disconnection is complete, we again enter HMOVE, in the opposite direction. When it is determined that reconnection needs to begin, we enter the RCON state.

When the node set has approached sufficiently close to the body, but before it regains contact, it enters the RWAIT state. This state's purpose is to accelerate the node set to match the speed of the body, before

contact has been achieved. It then waits for reconnection before entering HMOVE. During this state we also observe the adjacent node set's state; if it has entered the UPDATE state, then we will enter it as well.

Note that there is an opportunity to enter the UPDATE state, where the task level controller is consulted regarding the current setpoint. This opportunity occurs when enough nodes are in contact with the body to obtain an accurate determination of the body's location. If the body needs to be transported along a different heading, than the trajectory will come to a stop at this point and reinitialize for the new direction.

Let us assume that to maximize stability and comfort, we want as many nodes in contact with the body at a time as possible, N-a, where N is the number of node sets, and a is the number of disconnected node sets. Referring to Figure 8 for the dimensions, and using v_{XB} as the body velocity, and v_{XR} as the return velocity, and assuming instantaneous horizontal accelerations, we can derive the following equation for the period of the trajectory:

$$T = \frac{x_S}{v_{XB}} + \frac{x_S}{v_{XR}}$$
(16)

The time spent by the a node sets when disconnected or disconnecting from the body is given as Equation 17.

$$\frac{x_S}{v_{XR}} + \frac{x_S - x_B}{v_{XB}} = \frac{a}{N}T$$
(17)

Note that this is the portion of the trajectory shown between points 1 and 6 on Figure 8. Meanwhile, N-a node sets are in contact with the body:

$$\frac{x_B}{v_{XB}} = \frac{N-a}{N}T$$
(18)

Let us assume a contact fraction, α, which is the portion of the stroke x_S where the node set is in contact with the body:

$$x_B = \alpha x_S$$
(19)

Using this relation, and combining Equations 17 and 18 and solving for v_{XR}, we obtain the relationship between the forward body transfer speed v_{XB} and the return velocity v_{XR}:

$$v_{XR} = \frac{N-a}{a+\alpha N-N}v_{XB}$$
(20)

If α is near 1 and a is small, as we would require for comfortable motion, v_{XR} increases linearly with increasing N.

5 Body Position Control

The control of the node trajectories to generate desired motion could be referred to as low level control. High level control is required to give commands to the low level actuators, based on input from the user and accurate estimation of the body's location.

5.1 Closed-Loop Control of Body Position & Orientation

The system receives a setpoint from the user, containing information about the desired position and orientation of the body. This is compared with the estimation of the body's orientation and position (see section 4.2). Any error is passed along to the low-level actuators as a desired motion. A block diagram for this control is shown in Figure 10.
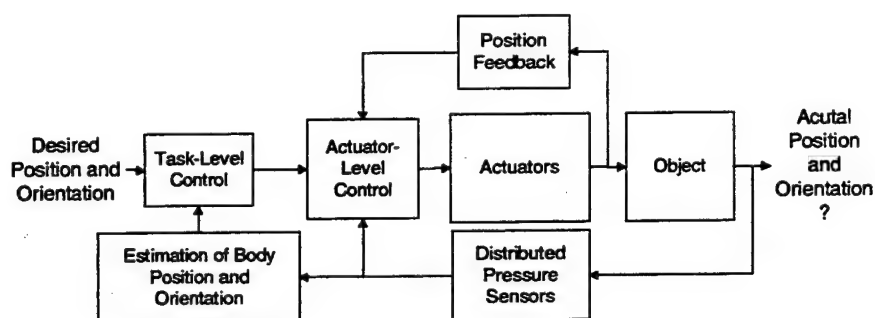

Figure 10: Control Block Diagram

5.2 Prediction Algorithms

The surface of the mattress is embedded with sensors to measure the pressure exerted by the human body. This information is used by the low level actuators to determine nodal contact/ non-contact information, but is also used by the high level actuators to determine the body's position and orientation.

When the system is first initialized, information about the body's shape, weight and position is either entered into the machine or obtained by careful sensor measurements. Once motion has begun, that information is used to estimate the current location of the body. Once every node trajectory cycle, the horizontal actuators report a vector of the body's displacement during that cycle. This is added to the current position estimate, and if correlated with the sensor data, becomes the new position estimate.

6 Experiments

Experiments were conducted using the prototype system to evaluate system performance.

6.1 Trajectory Evaluation

Experiments were conducted for both the square and trapezoidal trajectories. Several items of data were collected at every time step. The center of mass of the object as reported by the pressure sensors installed at the tip of individual nodes was collected, as was the position of each of the horizontal axes. The state of each of the node sets was recorded, as well as the state of contact of each node on the surface.

Below we compare the results of the same motion using the rectangular and trapezoidal trajectories. The data was obtained using the position estimation algorithm; the discontinuities in the data arise when the prediction is compared with actual sensor data, and if it does not fit, the object is moved to the apparent mass centroid, which may be inaccurate.
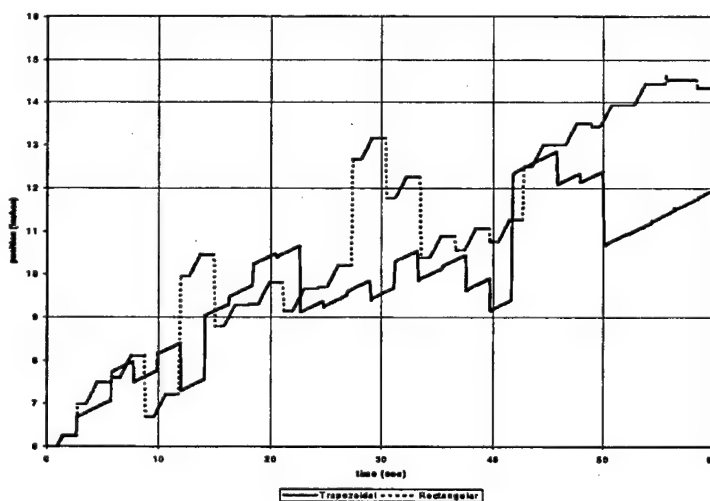

Figure 11: Comparison between Rectangular and Trapezoidal Trajectories

If we ignore the discontinuities, we note that the motion for the trapezoidal trajectory is slow but constant, while the rectangular trajectory produces more rapid motion in short bursts.

Let us examine the trapezoidal trajectory more closely. The first two plots below show the displacement of the body and the position offsets of the two horizontal axes; the third shows the discreet event states of the two axes with time. All three are for the time period between t=30 and 40 seconds.
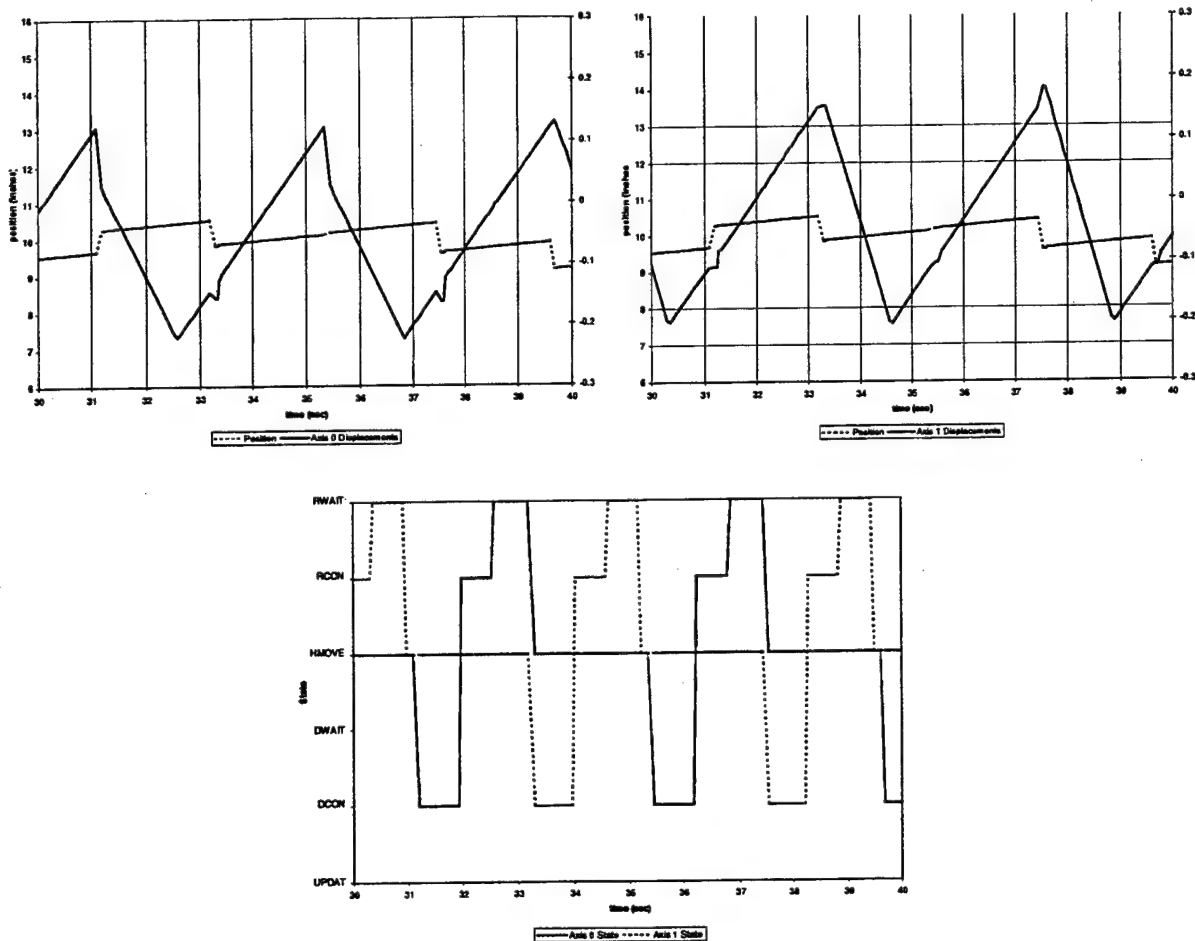
Figures 12: Plots of Axes 0 and 1

From these figures we can observe the operation of the system nodes. Note how at least one node set is in the HMOVE state at all times, with brief overlaps during transition. Also notice that the discontinuities in centroid position correspond to this overlap period, for this is when the estimate is correlated with the sensor data, and the loop is closed. (See Section 6.2). Observe that the DWAIT state is apparently skipped; this is due to a bug at the time of this writing.

6.2 Position Estimation Evaluation

In order to evaluate the performance of the closed loop position estimation algorithm, we must obtain position data from an outside source. Using a digital video camera, data points could be taken at set time intervals by visual inspection. A comparison for the rectangular trajectory is shown in Figure 13. The estimated data is parallel to the actual data, but has significant offset due to errors in the sensor measurement. The stair stepping phenomena is due to the quick motions during contact, followed by long pauses during reconnection.
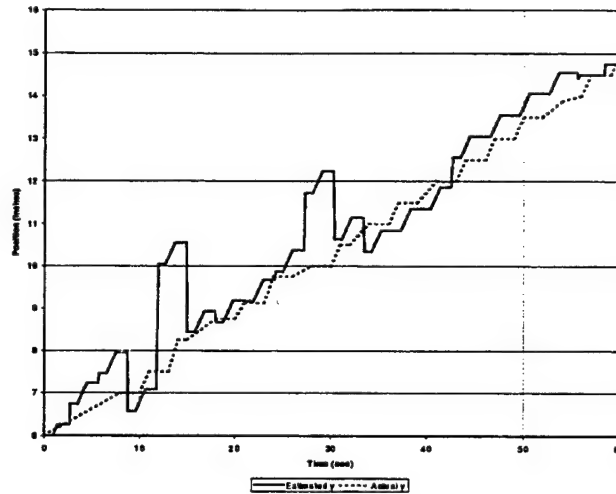
Figure 13: Comparison between Estimated and Actual Position, Rectangular Trajectory

The same procedure was followed for the trapezoidal trajectory in Figure 14. Again the estimated data is parallel to the actual data, but offset due to high sensor error. Eliminating this error will be a goal for future work.
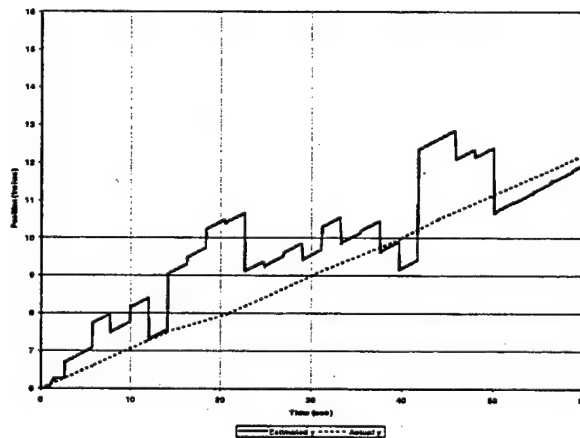

Figure 14: Comparison between Estimated and Actual Position, Trapezoidal Trajectory

7 Conclusion

An innovative system for the transport of bedridden patients has been described. This system has been developed to improve the quality of life for these patients and reduce the strain imposed on caregivers.

This work is the first to apply surface waves to the transport of human patients. Successful operation of the system depends on the design and control of the nodes which make up its surface. The choice of the trapezoidal trajectory was made to maximize patient comfort, reduce disconnection time to save energy, and to guarantee robust connections and disconnections despite actuator nonlinear behavior.

One the trajectory was chosen, coordination of the motions of the nodes in the trajectory had to be accomplished. The discrete event states used allow for errors and delays in the node set's motion.

The prototype that has been built to test these algorithms successfully moves objects across the surface, although it is too small to move a human. Translation and rotation have both been successfully demonstrated. A new prototype which makes use of pneumatics for the z-axis has been constructed, and should allow for more reliable motion at a higher speed.

References

[1] HFCA's Online Survey, Certification and Reporting Date of March 1997

[2] The Decubitous Foundation: Press Release. Feb, 1998. http://www.decubitus.org/ press/press.html

[3] Nonfatal occupational injury and illness incidence rates per 100 full time workers, by industry, 1996. OSHA. http://www.osha.gov/ oshastats/ bls/ Serv6.html

[4] Bureau of Labor Statistics. Number of nonfatal occupational injuries and illnesses involving days away from work, by occupation and selected parts of the body affected by injury or illness, 1994. Department of Labor, Office of Safety and health, 1996.

[5] Spano and Asada, "An Active Surface Wave Bed for Transporting Humans and Elastic Bodies", ASME IMECHE '98, Anaheim, CA November 15-21, 1998

[6] Mascaro, Spano, and Asada. "A Reconfigurable Holonomic Omnidirectional Mobile Bed with Unified Seating (RHOMBUS) for Bedridden Patients", IEEE Int. Conf. on Robotics and Automation, Albuquerque, New Mexico, April 1997.

Physical Aids

CHAPTER 14

A Semi-Active, Flexible, Beaded Support Surface for Tangential Transport of
Bedridden Patients
H. Asada, J. Spano

A Semi-Active, Flexible, Beaded Support Surface for Tangential Transport of Bedridden Patients

Haruhiko H. Asada
Principal Investigator

Joseph Spano
Graduate Research Assistant

Abstract

Bedridden patients and elastic bodies are transported by a novel ball transfer mechanism about their support surface. A series of spherical balls are constrained in their position in the support surface, but are allowed to rotate freely in three dimensions. A series of 'bed bugs' are coordinated to provide flexible support and tangential motion patterns of the human patient. This ball drive actuation allows a bedridden patient to move freely while lying comfortably upon the bed. First, the design concept is explained, then a description of the important parameters and performance measures is made. A mathematical model is formulated to describe the soft tissue interaction. Finite element methods are proposed to solve the mathematical model that take into account the non-linearity of the soft tissue constitutive law and the finite strain, large deformation formulation necessary to describe the physical phenomena. From these simulations technical requirements for transporting humans and elastic bodies will be obtained. A proof-of-concept prototype has been designed and will be tested to verify model results and the manipulation concept.

1.0 Introduction

The functional objectives of this project are to mechanically alter human posture in two dimensions and to provide flexible adjustment of support surface contour. The ability to carry out these two tasks has a large variety of applications in the service of humans. These include:

- Patient positioning to reduce joint stresses, increase circulation, provide massage therapy, and offer a positioning system for medical diagnostics such as MRI scanning.
- Patient transport for car ingress and egress and shifting the bed position.
- Orthopedic device emulation for neck positioning and spinal column positioning in rehabilitation environments.
- Active tissue therapy to promote maximum circulation in sensitive regions.
- A surgical tool for positioning the human body during surgery or to provide circulation in regions that may be cut off during long procedures such as open heart surgery.

Prior to this work two attempts have been made to achieve the goals of this project within the d'Arbeloff laboratory at MIT. [Spano and Asada, 1998] designed a mechanically generated one dimensional surface wave capable of moving humans. [Finger and Asada, 1999] have developed a two dimensional surface wave developed in part by shape memory alloy fibers capable of moving light weight rigid objects. The results of these efforts have been promising, but have required a complex mechanical system with limited flexibility. Two-dimensional surface waves have proven too complex with too many linkages and have faced difficulty with actuator limitations. The conclusion is that these systems are too complicated to achieve the manipulation objective with adequate nodal density to be clinically acceptable.

No other work in the literature has been detected where an attempt is made to study the mechanics or develop a system capable of moving the human tangentially across the support surface. Some work has been done in the study of the mechanics of skin and soft tissue. [Vannah 1996] has studied the material properties of bulk muscular tissue in vivo. [Fung 1993] discusses a variety of material models of the skin developed within the biomechanics community over the last two decades. The majority of the work in the

analysis area has focused on developing constitutive models of soft tissue with focus on skin, blood vessel, and heart tissue. Work in the application area of soft tissues support systems has focused on the design of custom contoured cushions and prostheses as a means of reducing peak pressure at the support surface interface. [Brienza 1996] details a test apparatus useful for this purpose.

In this paper a new tack on the problem is undertaken, inspired by the use of spherical support surfaces that are in use in seats that support people for long periods of time such as taxi drivers and long-haul highway truck drivers. Many individuals use a beaded support surface or a lumpy support surface molded into the foam of the chair to allow for better air circulation and stimulation of the tissues. This promotes skin health in situations where the body is in contact with the support surface for long periods of time. A step beyond this is proposed where we activate these spherical support units and use them as a tangential transport mechanism for bedridden patients.

2.0 Concept

As mentioned before, the surface wave actuator concepts previously explored showed extremely high complexity as you increase the nodal density of the support surface. However, our newest design can realize acceptable densities with little change in mechanical complexity and this indeed is the true benefit of this concept.

Our new manipulation concept is termed the MIT mattress. The mattress possesses two components: a semi-active sheet and active 'bed bugs'. Working together the 'bed bugs' cooperatively reconfigure the human lying on top of the active sheet surface. The overall system components are shown in Figure 1. The semi-active sheet is a matrix of spherical elements that are free to rotate in the three principle directions, but each sphere is constrained in its position. See Figure 2. Below the spherical foundation matrix a set of holonomic, omni-directional vehicles move in a plane parallel to the support surface. Each vehicle features a conveyor belt that can be precisely positioned beneath the support surface. A vertical degree of freedom on the vehicle engages the conveyor belt with the appropriate set of spheres. When the belt is actuated frictional contact between the belt and the sphere, drives the sphere causing it to spin in the direction set by the vehicle. This

3

spinning motion causes the body lying on top of the surface to be translated.
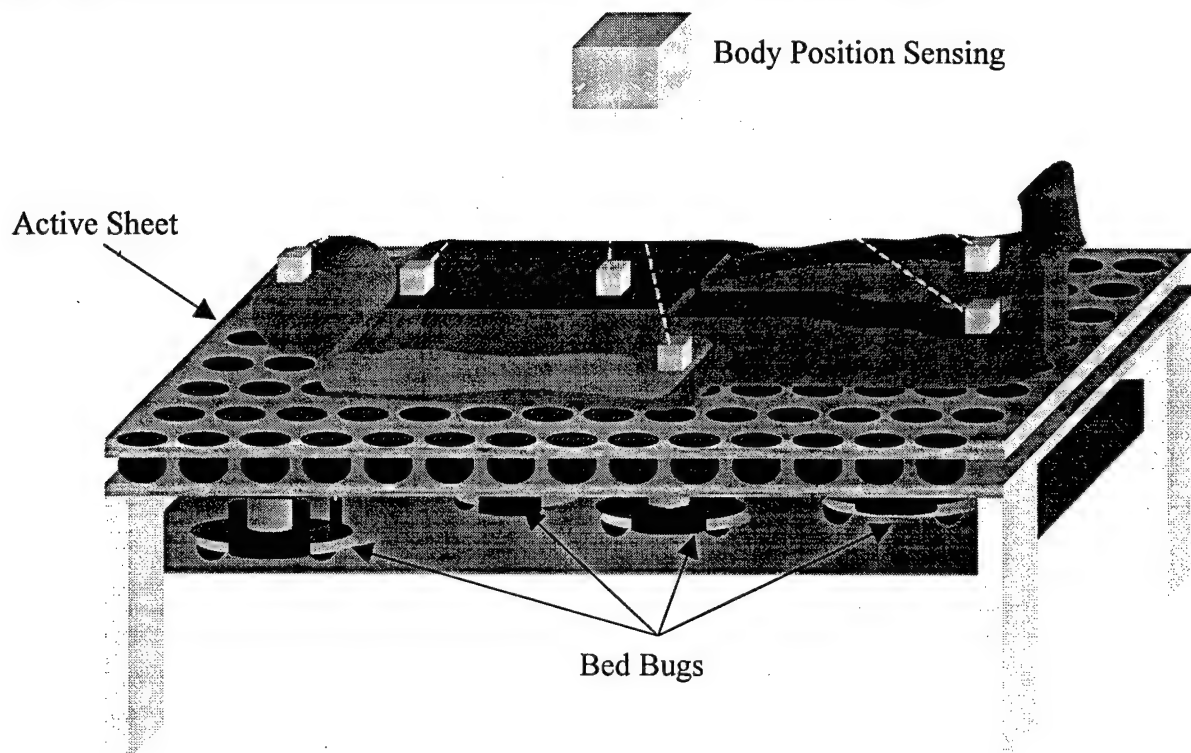

Figure 1 - Semi-active, beaded support surface

The active sheet is supportive like a conventional mattress. Functionally, the active sheet offers two modes of support. The first mode is a resting mode that offers support comparable to a typical support surface. In addition in active mode the body is supported by spherical balls for tangential transport.
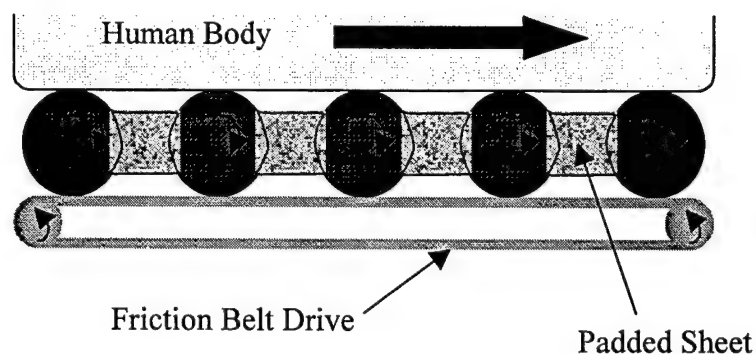

Figure 2 - Drive mechanism detail

The 'bed bugs' offer self-contained, holonomic, omni-directional motion in the bed plane. A vertical degree of freedom engages the active sheet and puts it into active mode.

4

A belt driven friction drive system drives the active sheet and moves the human tangential to the support surface.
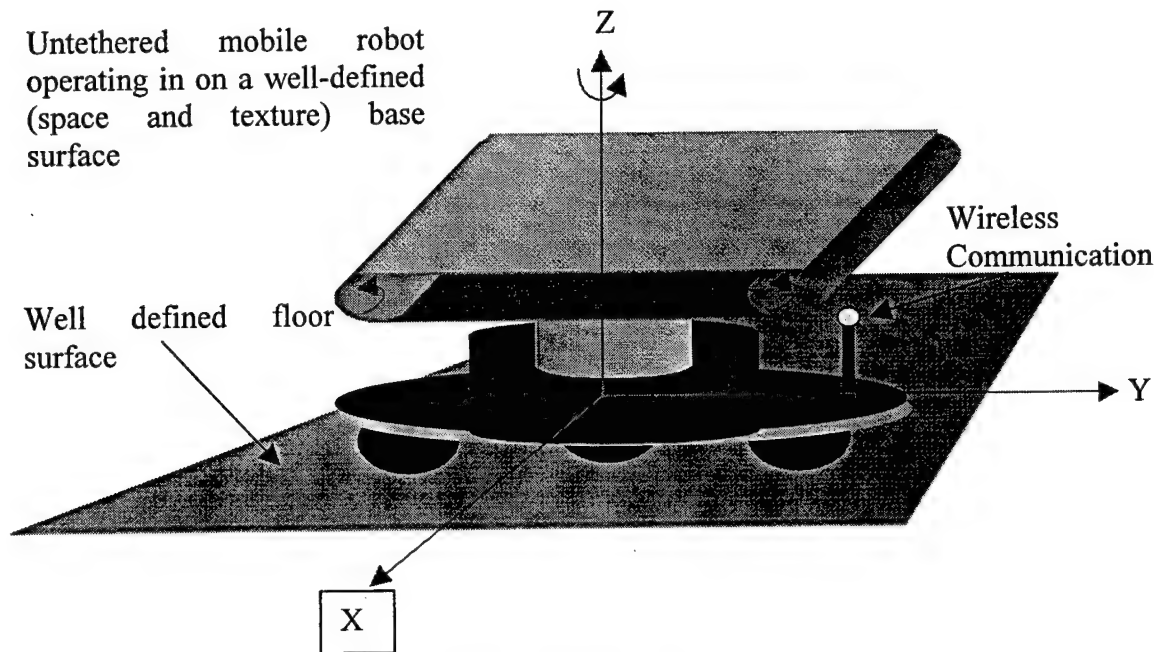

Figure 3 - 'Bed bug' detail

The net result is an active mattress that mimics the comfort level of a conventional mattress, but also allows for flexible manipulation of the human body in the plane of the support surface.

Semi-active systems have several interesting properties. The bed surface may not be active at all times. Instead each segment may be activated one by one sequentially. Since for our application it is not necessary to activate all support units independently and simultaneously there is no need for dedicated distributed actuators for each individual segment of the support surface. Furthermore, it is very desirable to reduce the number of actuators and simplify the design as much as possible to keep costs of production to a minimum. The result is a centralized, few degree of freedom actuator system that is engaged with a selected portion of the semi-active bed surface in the appropriate direction when, and only when it is needed.

This semi-active system design features omni-directional, continuous transport. Some of the salient features of this proposed method are the fact that motion is continuous, not intermittent or disjoint. This results in the smoothest possible manipulation of the human

5

body with a minimum of extraneous stresses at the interface resulting from repositioning of the support mechanism. From the design perspective the omni-directional transport system is a relatively simple mechanism that is very reliable.

Control of the semi-active system involves the coordination of the 'bed bugs' to achieve a particular posture reconfiguration objective. These objectives will be established by a human body position planner that will prioritize manipulation activities. First priority will be to shift the support points of the active surface to ensure that no point on the body is subjected to high pressure for an unacceptable length of time. Other priorities need to be designed by the medical staff to customize the treatment program for each individual patient. Example treatments are given in the application examples at the beginning of the report.
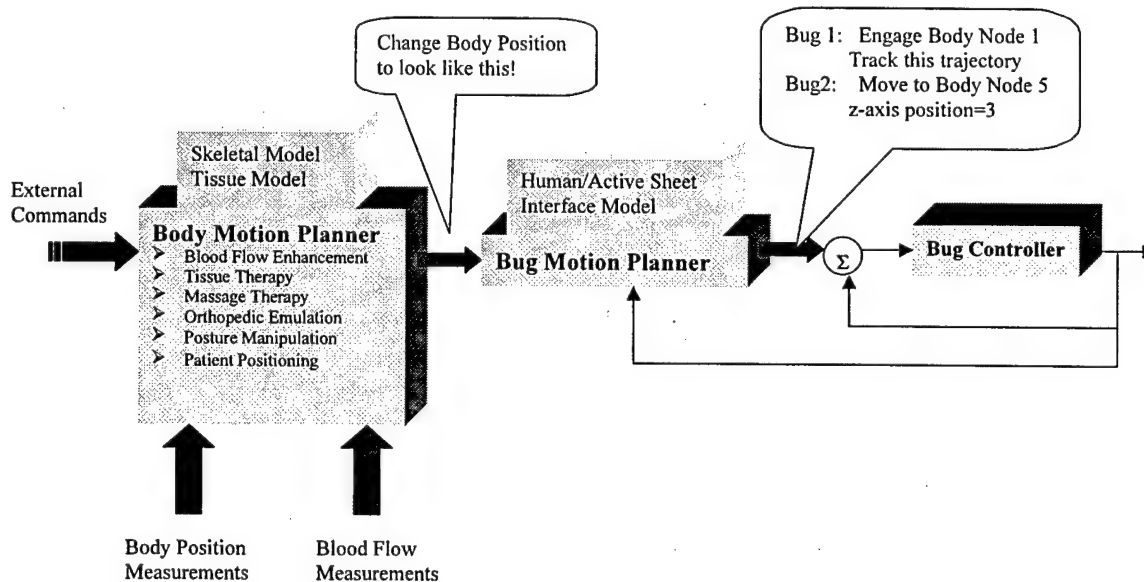

Figure 4 - Control architecture

In addition to the design issues there are issues of control and sensing and issues of planning and cooperation to realize the complete system. For this initial project work we will focus on the analysis of the support surface itself without regard for the 'bed bugs'. Once this support surface is understood and its promise as a meaningful manipulation prospect for bedridden patients proven we can move forward on these other issues. Mechanical analysis of the human/active mattress interface is important to establish the design conditions for successful manipulation. We must establish criteria for moving non-

homogeneous, elastic bodies on spherical foundation in a clinically acceptable manner. This work will require use of FEA to handle the anisotropic constitutive behavior of the human and the finite strain formulation needed to accurately describe the deformation that occurs at the human/active mattress interface.

The friction law between the human and the active mattress must be investigated to reveal what materials are required to provide adequate tractive forces required to transport the human across the support surface.

3.0 Mechanical Analysis of Human/Support Surface Interface

The purpose of this analysis is to clearly formulate the mathematical problem of supporting soft human tissue upon a spherical foundation. The finite element solution of this mathematical problem will then provide insight into the design of the support surface. Specifically, physiologically accepted values for maximum acceptable interface pressure must be established so that the system can be designed such that adequate support can be provided. The questions that must be answered are:

- ✓ Is the device capable of meeting its functional objective of tangential manipulation?
- ✓ Is the device clinically viable by offering a physiologically acceptable (healthy) support surface?

3.1 Physical Characteristics that must be Modeled

Before beginning the mathematical problem formulation for the contact scenario that we wish to model, we must identify the physical characteristics of the system that comprise the system. The spherical foundation matrix model is transparent. The geometry is well defined and the spheres are assumed to be rigid bodies. The difficulty lies in the modeling of the human side of the interaction.
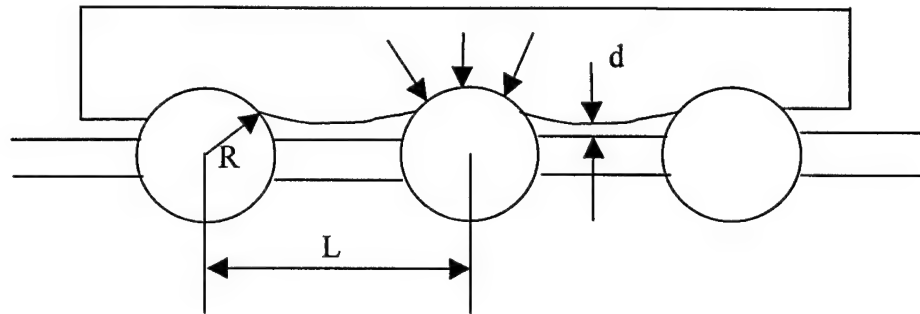
Figure 5 - Schematic of human/support surface interaction

From Figure 5 one can see that the deformation of the soft tissues is significant. So already it is clear that an accurate model must include strain-displacement definitions that will be able to account for the large deformations and large strains taking place in this hyperelastic material. To further complicate matters the constitutive relations of soft human tissues in vivo are extremely complex. They exhibit anisotropic, non-linear, viscoelastic behavior. Modeling all of these features explicitly has not been done. However, the appropriate combination of these features can be approximately modeled and it is important for us to identify which features are the most important to capture. [Vannah 1996] describes experimental results that indicate that stress relaxation is completed in approximately one second after the application of the load. For our system, this one second is irrelevant, therefore the viscoelastic nature of the material does not need to be specifically accounted for in the constitutive model. This leaves us analyzing a hyperelastic, non-linear, anisotropic material. Methods exist to model hyperelastic, non-linear, anisotropic materials in a mathematically rigorous way that is consistent with thermodynamic principles. However, at some point experiments must be conducted to determine material constants no matter what formulation is used, and if one treats the full anisotropic case these constants will have to be obtained. One observation that has been made is that soft tissues have a plane of elastic symmetry tangential to the skin surface. A second observation is that in the case of a human body supported on a spherical foundation matrix, the loading will always be coming from the same direction. So in essence one can ignore the anisotropic aspect of the problem because the soft tissue will never be loaded in a manner other than normal to the skin surface.

8

This leaves us with a finite displacement, finite strain, hyperelastic, non-linear, isotropic formulation. This conjecture is supported by experimental results by [Vannah 1996].

3.2 Schematic Model of the Real Physical Situation

Analysis should relate a variety of design parameters for this system shown in this schematic model.
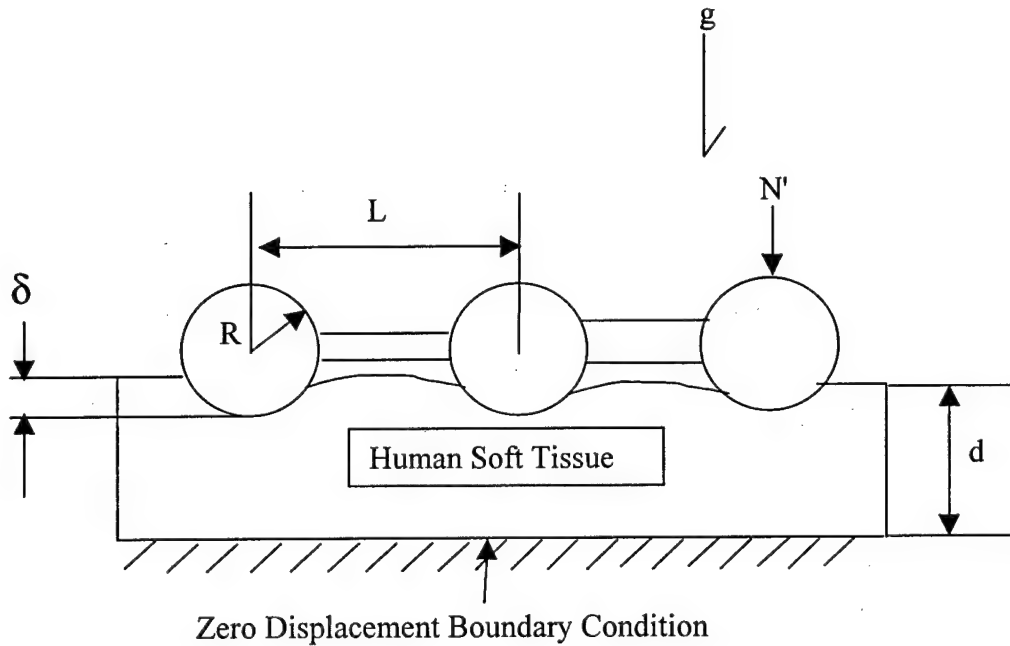

Figure 6 - Schematic of Mathematical Model

The parameters of this mathematical representation of the real system are defined as follows:

d-depth of soft tissue

L-space between adjacent support spheres

R-radius of support spheres

N'-Effective reaction load on each sphere to cancel gravity load of human body

δ – Depth of impingement of support spheres

Modeling assumptions are as follows:

The support spheres are modeled as rigid bodies of radius R. They are constrained to move only in the direction normal to the soft tissue surface. The soft tissue is characterized by a depth, d. The soft tissue is characterized by a zero displacement

boundary condition at this depth, d. The reason for this modeling assumption is that we would like to characterize the stresses induced when the skin in pinched between the support surface and bone. A review of the literature indicates that a zero displacement boundary condition between soft tissue and bone does occur in the body and the assumption here is that the bones are rigid and fixed. The impingement distance of the support sphere is labeled as δ. This distance is a consequence of the support scenario, but a requirement for adequate system performance is that this distance is not so great that the body drags on the spherical support retaining layer and impedes the body motion by creating drag forces. Finally the load on each sphere is represented by N'. The most important aspect of this problem is the material law assumption for soft tissue. As we discussed in the previous section we would like to capture the non-linear, finite displacement, finite stain, hyperelastic behavior of the tissue. In the linear case the Mooney-Rivlin formulation for the strain energy function is often used to analyze these types of 'rubber elasticity' problems, but due to the non-linear nature of the tissue [Vannah 1996] is able to fit data reasonably well using a Jamus-Green-Simpson strain energy function of the form

$$ W = c_{10}(I_1 - 3) + c_{01}(I_2 - 3) + c_{11}(I_1 - 3)(I_2 - 3) + c_{20}(I_1 - 3)^2 + c_{30}(I_1 - 3)^2 $$

where the invariants, expressed in terms of stretch ratios are:

$$ I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 $$

$$ I_2 = \lambda_1^2\lambda_2^2 + \lambda_2^2\lambda_3^2 + \lambda_3^2\lambda_1^2 $$

For a volume conserving material $I_3 = 1$. [Vannah 1996] gives experimentally obtained values for the parameters of the strain energy function and the formulation is complete.

3.3 Performance Measures

It is important when establishing performance measures to establish specifically what service we want to offer the bedridden patient. In our work we would like to see the health of the bedridden patient assume the top priority. Our design concept has the opportunity to offer a support surface that can regulate the posture and position of the human and the question remains what is the healthiest configuration in which to manipulate and support the human. Health in this context can be defined as providing an

acceptable level of blood and air circulation to all of the tissues at the human/support surface interface. It should be emphasized that a healthy support surface may not always be the plushest, comfortable support surface. For instance, in the case of people in occupations with long periods of sitting, often a beaded mat is used between themselves and the seat. This mat may not be as purely comfortable as a cushioned seat, however the added benefits of increased air circulation far offset this discomfort and results in a situation which is healthier.

When an individual lies upon a support surface stresses are induced which deform the soft tissue. The changes in shape result in occlusion of blood vessels and lymphatics, and stimulate nerve endings which could signal discomfort to the central nervous system. In addition to the deformation of the tissues there is an exchange of heat between the patient and the surface, so thermal considerations come into play also in the physical interaction. It should be emphasized in this work that psychophysical and thermal responses will not be treated as criteria for constraining the design. The attempt is made in this work to define healthy support surface conditions in a rigorous, quantified manner that does not rely on psychophysical responses. The reason for this is that these responses are based on nervous system phenomena that are not remotely understood by the engineering or medical community. In addition, for the problem addressed by this work the use of psychophysical responses is not seen to be necessary to define healthy support surface design. In other words we would like to show that one only has to appeal to physically understood mechanical phenomena such as heat transfer, fluid flow, and solid mechanics to properly analyze the human-support surface interface and synthesize the appropriate geometry to ensure a healthy interaction. With respect to thermal interactions the current design concept does not attempt to control temperature at the interface, however, the overall shape of the support surface will affect heat transfer between the patient and support surface. These effects will not be analyzed rigorously, nor included in the design constraints. [Ferguson-Pell 1990] claims that sweating can become a serious issue if the interface temperature exceeds 38 degrees centigrade.

Important measures of performance from the perspective of *machine* performance include the height the body lies above the nominal support surface. In addition the

11

maximum traction force which corresponds to the system's ability to move the human efficiently with a minimum amount of slippage is also important.

From the *human* perspective there are a number of issues of importance. To categorize and prioritize these issues [Ferguson-Pell 1990] has written a primer on the clinical criteria of seat cushion selection. In his report the most important factor that determines <u>functionality</u>, not just comfort, of a support surface is the proper distribution of stress in soft tissues, followed by control of moisture accumulation and heat. Clearly from a clinical and physiological perspective these are the criteria that all support surface designs must meet to be acceptable. This report is a general overview and more detailed quantified data is needed for a rigorous mechanical analysis and design.

The literature in the fields of physical medicine, rehabilitation, and biomechanics offer a wealth of quantitative information regarding human physiology. Although this information is often collected with no specific application in mind or for applications unrelated to this project, some facts are relevant to this work with some interpretation.

Of primary importance is the data showing the relationship between allowable pressure and time duration for 'safe' pressure-time support. This data illustrates clearly the maximum allowable pressure for a given period of time. What this data demonstrates is that below some pressure threshold, tissue can sustain itself without damage for an infinite amount of time. This is the region where fluidized beds or special contoured surfaces are designed. These surfaces are used in the treatment of bedsores by keeping pressure as evenly distributed as possible. However, this data also indicates that some higher pressures, above the threshold can be maintained for finite periods of time. This becomes important when one considers that our surface is semi-active and could conceivably reconfigure to redistribute pressure periodically. In this scenario, although peak pressure could be above the threshold permanent damage could easily be avoided.
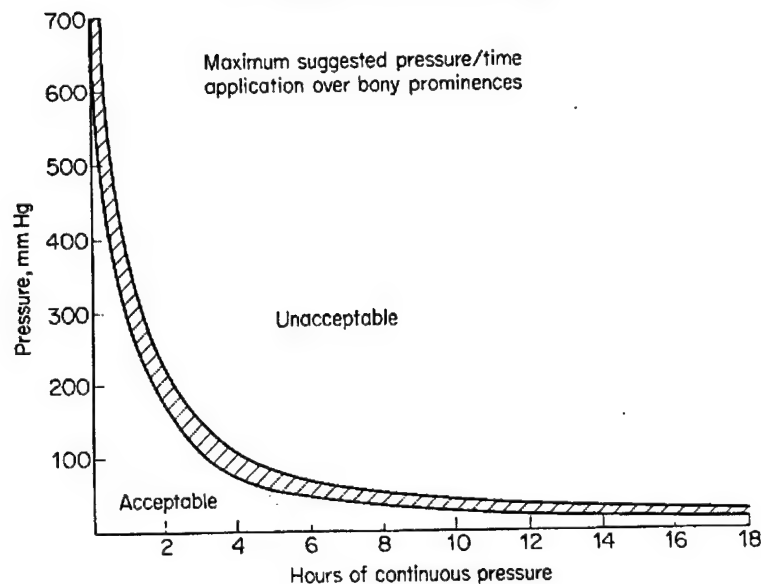
Figure 4. Allowable pressure vs. time of application for tissue under bony prominences. Curve gives general 'guidelines' and should not be taken as absolute

Figure 7 - Pressure-time data

The data presented by Reswick is a clinically proven threshold for normal pressures. However, in a real support surface-human interaction, shear stresses will also be induced and it is important to note the influence of these shear stresses on the human. [Bennett 1979] presents experimental work on blood flow occlusion under shear load. His experimental results show that shear is a contributing factor to blood flow occlusion, but not as dramatically as normal stress. In general it was concluded that normal stress is about twice as effective at producing blood flow occlusion. [Guttmann 1979] makes it clear however that shear stress situations must be avoided, citing a dramatic examples of plaster casts to treat traumatic paraplegics. It was thought that these rigid casts would distribute pressure very evenly, however, in fact over time, changes in shape of the human body created high shearing actions that produced terrible bedsores. The moral of the story is that both shear and normal stresses must be carefully controlled and the guidelines set by [Reswick 1975] are a clinically dependable, quantified constraint on our design.

Another piece of information that is important is the threshold of pain for tissue under pressure. Pressure algometry (dolorimetry) has been used to evaluate sensitivity to pain and the assessment of pressure perception. Pressure threshold is defined as the minimum

13

pressure which induces pain or discomfort. A survey of a number of works in the field indicate a range of .2-.9 Mpa as the threshold of pain for normal people. [Fischer 1987] The variation depends on the individual person and with specific trends on the position where the measurement is taken. Pain is certainly a psychophysical response based heavily on the nervous system response of the human. These types of responses are not understood in a rigorous way and use of this data in the design of our support surface must be approached with caution. This data will be used as an upper bound on designs. In other words, designs that cause stresses of this level will be discarded due to their likelihood to be painful and therefore clinically unacceptable. No other use of this data, particularly in our mathematical analysis will be made.

4.0 Testbed Implementation

4.1 Objective

Our objectives for testbed implementation are two-fold. First we would like to establish that given clearance between the human body and the spherical retaining layer, the human can be transported by the proposed belt-driven, ball-transfer mechanism. Preliminary experiments in lab have shown that the body can be supported above the plane of retainers using 1" diameter balls and an internodal distance of 2". As a sample, worst-case manipulation scenario, we have designed an experiment that is likely to be the most difficult manipulation task that will be faced by the bedridden patient. That task is the movement of the human hip position across the support surface. In this scenario the concentration of load is higher than any other scenario because all of the human body weight is being supported by over a relatively small area. We would like to show that even under these extreme conditions tangential motion can still be achieved.

Second we would like to confirm the results of our finite element analysis of the tissue/support surface interaction by comparing the deformations predicted by the finite element model and those measured on the testbed to see that indeed the model is adequate to capture the mechanical behavior.

4.2 Design

The experimental prototype includes a fixed belt, fixed socket, working prototype prototype that is capable of moving the hip position.
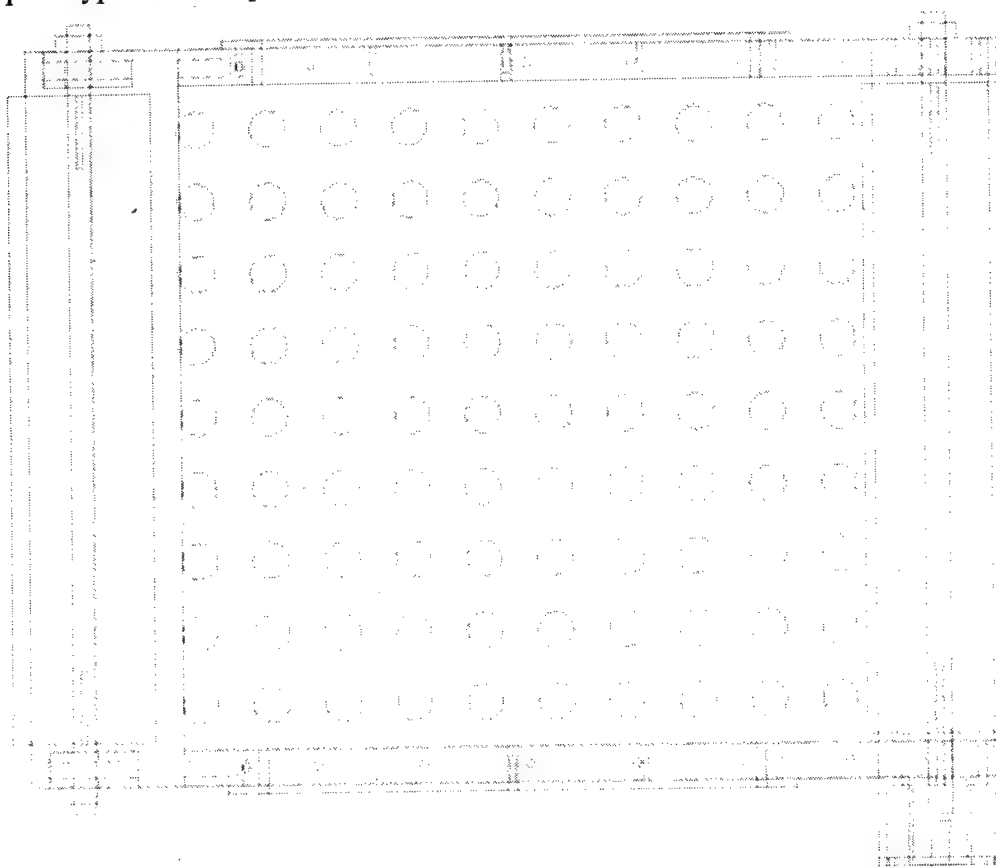
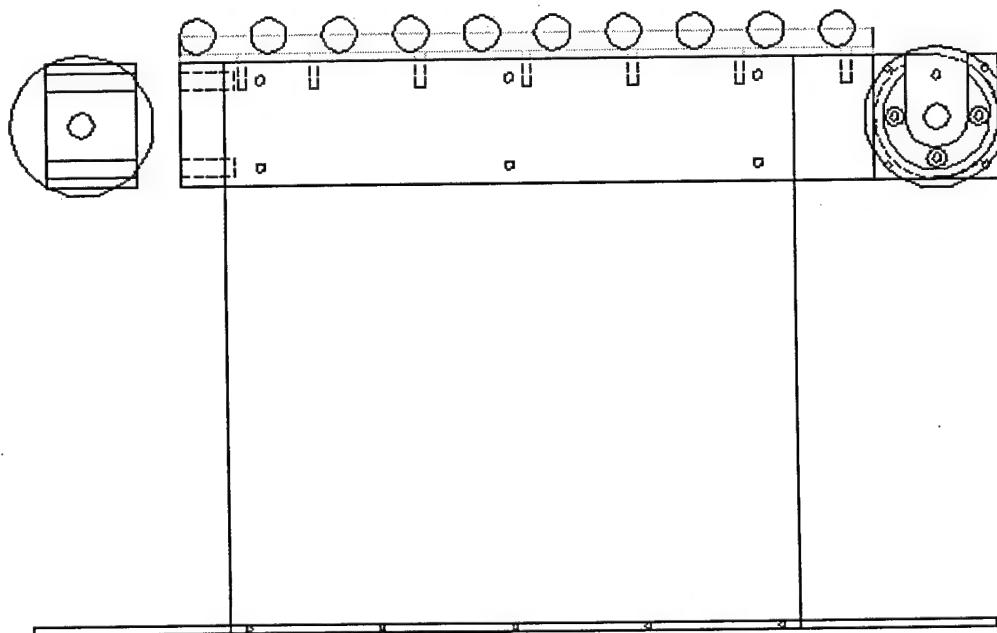

Figure 8 - Top View of Testbed

Figure 9 - Side View of Testbed

5.0 Conclusion

In this work we have established a design concept for mechanical manipulation of the human tangentially on the bed surface. However, we recognize that this novel support surface must be carefully designed to ensure that the physiological health of the bedridden patient is not compromised. Specifically we show the rigorous formulation of the mathematical problem associated with supporting soft tissue on a spherical foundation matrix. By finite element solution of this problem we want to show that it is possible to design a support surface of this nature that will impose stresses on the human tissue that are below established levels required for permanent damage of tissue due to blood flow occlusion. The solution of this model will have to be tested by experiments before a final design is established. This final design will be a clinically acceptable support surface capable of planar manipulation by our semi-active system concept.

[Spano and Asada, 1998] J. Spano and H.H. Asada, "An Active, Surface Wave Bed for Transporting Humans and Elastic Bodies," ASME IMECHE '98, Anaheim, CA November 15-21, 1998

[Finger and Asada, 1999] W. Finger and H.H. Asada, "Design and Control of an Active Mattress for Moving Bedridden Patients", IEEE Robotics and Automation Conference, Detroit, MI April 1999

[Brienza, 1996] Brienza, D.M., et.al., "A system for the analysis of seat support surfaces using surface shape control and simultaneous measurement of applied pressures." in *IEEE Transactions on Rehabilitation Engineering*, pp103-113, June 1996.

[Fung, 1993] Y.C. Fung, Biomechanics: Mechanical Properties of Living Tissue, Springer-Verlag, Second edition, 1993

[Vannah, 1996] William Vannah and Dudley S. Childress, "Indentor tests and finite element modeling of bulk muscular tissue in vivo," *in Journal of Rehabilitation Research and Development*, Vol. 33 No. 3, July 1996 pgs. 239-252.

[Fischer 1987] Andrew A. Fischer, "Pressure algometry over normal muscles. Standard values, validity and reproducibility of pressure threshold", in *Pain*, 30(1987) 115-126.

[Ferguson-Pell 1990] Martin W. Ferguson-Pell, "Seat Cushion Selection", *Journal of Rehabilitation Research and Development Clinical Supplement*, vol.2, pp. 49-74, 1990

[Reswick 1975] J.B. Reswick and J.E. Rogers, "Experience at Rancho Los Amigos hospital with devices and techniques to prevent pressure sores", Bedsore Biomechanics, 1975

[Bennett 1979] Leon Bennett, et.al., "Shear vs. pressure as causative factors in skin blood flow occlusion", *Archives of Physical Medicine and Rehabilitation*, Vol. 60, July 1979, pp.309-314

[Guttmann 1975] Sir Ludwig Guttmann, "The prevention and treatment of pressure sores", <u>Bedsore Biomechanics</u>, 1975

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Human-Machine Interface

CHAPTER 15

Human Machine-Interface and Interactive Control
Part 1: Photo-Plethysmograph Nail Sensors for Measuring Finger Forces without
Haptic Obstruction: Modeling and Instrumentation
H. Asada, S. Mascaro, K-W Chang

Human-Machine Interface and Interactive Control

Part 1: Photo-Plethysmograph Nail Sensors for Measuring Finger Forces without Haptic Obstruction: Modeling and Instrumentation

H. Harry Asada
Professor, Principal Investigator

Stephen Mascaro
Graduate Research Assistant

Kuo-Wei Chang
Senior Lecturer

Abstract

A new type of touch sensor for detecting contact pressure at human fingertips is presented. Fingernails are instrumented with arrays of miniature LEDs and photodetectors in order to measure changes in the nail color pattern when the fingers are pressed against a surface. Unlike traditional electronic gloves, in which sensor pads are placed between the fingers and the environment surface, this new sensor allows the fingers to directly contact the environment without obstructing the human's natural haptic senses. The finger force is detected by measuring changes in the nail color; hence the sensor is mounted on the nail side rather than the finger pad. Photo-reflective plethysmography is used for measuring the nail color. Hemodynamic modeling is used to investigate the dynamics of the change in blood volume under the fingernail. Miniaturization techniques are presented and implemented on a prototype sensor. Applications to human-machine interaction are discussed.

1. Introduction

Electronic gloves have been extensively studied in the past decade in the robotics and virtual reality communities [1]. There are many ways of providing force feedback to the human from a virtual environment or from sensors on the robot, and many electronic gloves now make use of such force feedback [2][3][4]. A good example is the CyberGlove developed by Kramer [5]. However, few electronic gloves collect touch-force data from the human fingers as the human interacts with the environment. The ones that do make use of pressure sensing pads consisting of conductive rubber, capacitive sensors, optical detectors, or other devices which are placed between the fingers and the environment surface [6][7][8][9]. These sensor pads, however, inevitably deteriorate the human haptic sense, since the fingers cannot directly touch the environment surface.

In this paper, a new[11] approach to the detection of finger forces is presented in order to eliminate the impediment for the natural haptic sense. Namely, the finger force is measured without having to place any sensor pad between the finger skin and the environment surface, but is detected by an optical sensor mounted on the fingernail. This allows the human to touch the environment with bare fingers and perform fine, delicate tasks using the full range of haptic sense. In previous work, the basic principle of such a sensor is described and an initial prototype is implemented [10].

Although the original prototype was successful in measuring the gross change in color of the fingernail caused by touch force, it has two limitations. Firstly, by using a single LED of only one wavelength, the sensor is not able to give complete information regarding the state of the finger at any particular point. The three factors that determine the color at any point are the 2 hemoglobin concentrations, as well as the volume of blood. By using only one wavelength of light, one cannot get much insight into the nature of the color change that is measured. Secondly, by measuring the color change at only one point, the sensor is not able to give complete information regarding the cause of the color change. One finds that certain color changes can be

---

[1] See appendix on patent search

2

attributed to bending of the finger rather than touching a surface. This is a serious cause of noise and false alarm when monitoring the fingernail for touch forces.

The goal of this paper is to better understand the physiological behavior behind the color change through modeling and experimentation, and to obtain design guidelines and signal processing algorithms for the nail sensors. The nail can then be instrumented with arrays of multiple LEDs and photodetectors to perform the correct measurements.

Glove based input has been used increasingly in the last decade for teleoperation and other forms of human-machine interaction. Sturman and Zeltzer provide a comprehensive review of glove-based input [1], including applications to teleoperation and robotic control. Postural gesture recognition has been applied to teleoperation of robots and to teaching of robots by demonstration and guiding [11][12][13]. Recently, Voyles and Khosla developed a system for teaching-by-guiding by inferring human intentions from tactile gestures, which were measured by force sensors on the robot [14]. However, such as system is not very flexible, as it requires modification of the hardware on the robot. Instead, by measuring touch forces on the human side, we can achieve a much greater flexibility in measurement. At the end of this paper, we will present several applications that take advantage of the unique characteristics of the new touch sensor.

2. Principle

2.1 Hemodynamics of Fingers and Nails

As a finger is pressed horizontally on a surface with increasing force, a sequence of color changes is observed through the fingernail. In fact, the color change is characteristically non-uniform across nail, resulting in distinct patterns of color change. Figure 1 shows a typical sequence of noticeable color changes with increasing force. The force values shown are derived from one particular person. Although they may vary from person to person, the underlying physiological principle is universally applicable for the healthy fingernail in an ordinary environment.
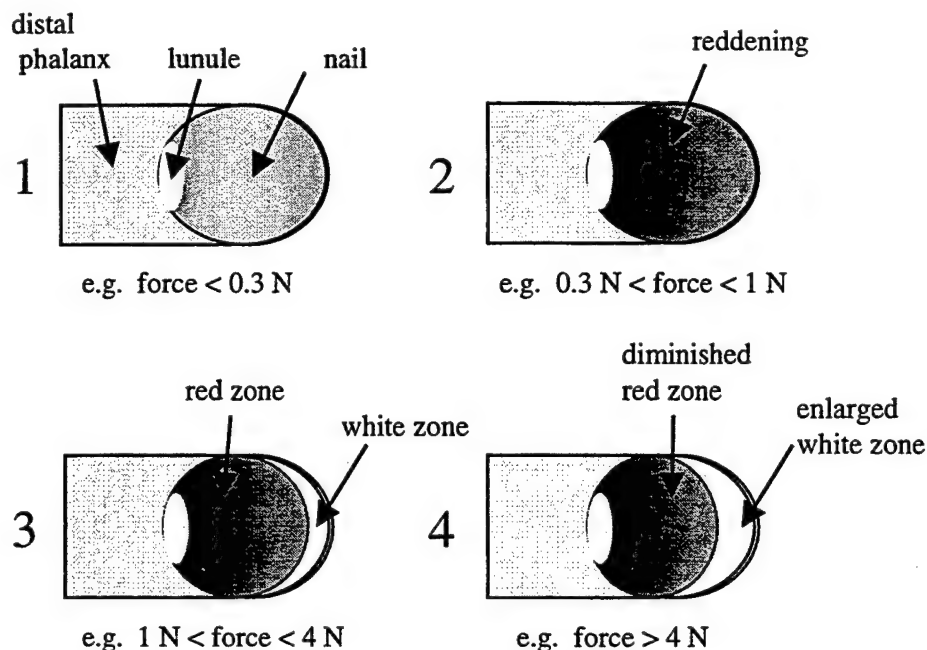
3

Figure 1: Typical Fingernail Color Changes

Above a certain threshold, e.g. 0.3 N, the entire nail begins to redden in color. Now, the dermis is richly vascularized with large arteriovenous shunts [15]. Directionally the capillaries are vertical into the dermal papillae under the nail matrix, but longitudinal under the nail bed [16]. Forces within the 0.3 N to 1 N range are sufficient to cause the venous return of blood in the fingertip to be progressively constricted. This results in pooling of arterial blood in the capillaries underneath the fingernail. This arterial blood is rich in oxy-hemoglobin and therefore bright red in color. When the contact pressure reaches a point, e.g. 1 N, the vein is completely blocked and the fingernail color stops reddening with further increase in touch pressure.
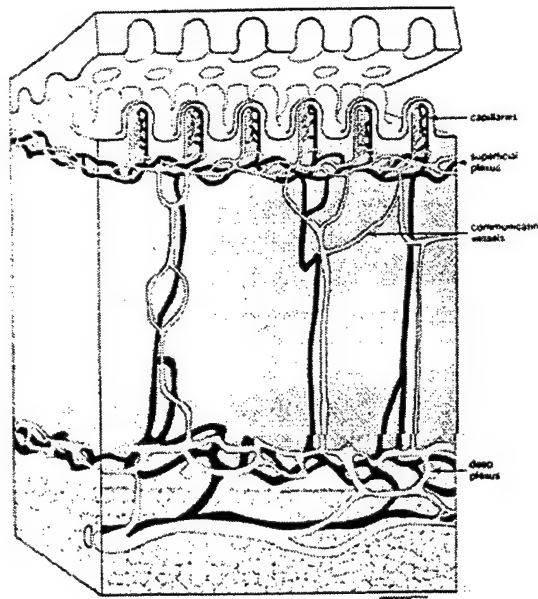
Figure 2: Dermal Vasculature (From Moschelle, S.L. and Hurly, H.J.: Dermatology, Vol. 1, pp. 35, Saunders, 1985)

Now further increases in touch pressure begin to constrict the arterial supply at the tip of the finger, causing the blood to be pushed out of this region, resulting in a white band at the tip of the finger. The rest of the nail remains deep red, as the capillaries are protected from the pressure of the touch force by the bone of the distal phalange, which is connected to the fingernail via a strong matrix of collagen and elastic fibers. The nails are thus described as "immobile over the distal phalange" [16]. In fact, the fingernails have an important tactile function in providing "support and counterpressure for the digital pad, thereby aiding manipulation" [15].

As the force increases, the white band widens until some limit is reached, e.g. at 4 N. Further increases in touch force beyond this point have no visible effect. However, forces applied longitudinally to the front of the fingertip and shear forces along the same direction are more effective at exerting stresses on the tissue above the bone, and are capable of increasing the white band even further.

This phenomenon can be utilized to measure touching force and contact pressure by monitoring changes in fingernail color without having to put a sensor between the finger and

surface. The change in color is directly related to the pooling of arterial blood and its oxy-hemoglobin saturation (relative concentrations of oxy- and reduced- hemoglobin). The amount of blood and the oxygen (oxy-hemoglobin) saturation under the fingernail bed can be monitored by shining light into the fingernail and measuring the reflectance.

2.2 Construction of the Nail Sensors

An example of the experimental setup is shown in Figure 3. A red LED at 660nm illuminates the nail bed with a red light. A photo-transistor is mounted on one side of the LED and catches the reflected light from the nail bed. As contact pressure increases, more arterial blood accumulates under the nail, resulting in two competing phenomena. On one hand, the additional volume of blood under the nail increases the effective path length over which the light is absorbed, tending to increase the absorption of light. On the other hand, the increase in oxygen saturation of the blood under the fingernail decreases the absorption coefficient of the blood, tending the decrease the absorption of light. In initial experiments [10], the first phenomena turned out to be the dominant of the two. As the contact pressure increases, more light is absorbed, less red light is reflected, the impedance of the photo-transistor drops, and the output voltage, Vout, increases. Vout reaches an asymptotic value when the veins are collapsed and closed shut.

If an infrared LED at 940nm is used instead, the output voltage will increase at a larger rate. This is because the trends of the absorption curves for hemoglobin and oxy-hemoglobin reverses after crossing the isobestic point at 800nm (Figure 4), and an increase in oxygen saturation will result in a larger coefficient of absorption. With high contact pressure, the increase in effective path length and the increase in coefficient of absorption will work together to increase the absorption and decrease the reflected IR light at 940nm.
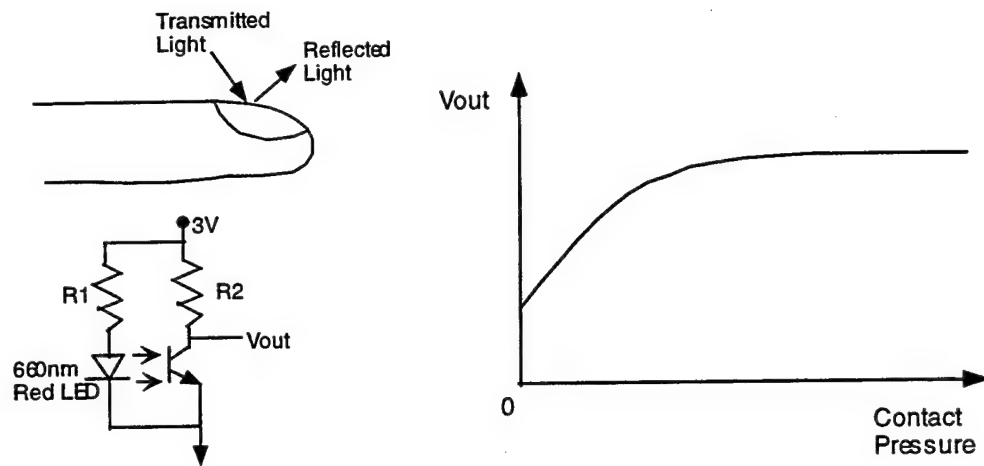
Figure 3: Plethysmograph Fingernail Sensor

The red LED at 660nm and the IR LED at 940nm can be used in the same sensor to measure relative concentration of oxygen in the blood. The two types of LEDs are illuminated alternately and the reflected lights are measured by the same photo-detector with the aid of sample-and-hold circuitry. Use of a third LED at the isobestic wavelength can provide measurements of absorption that are decoupled from the effect of oxygen concentration.
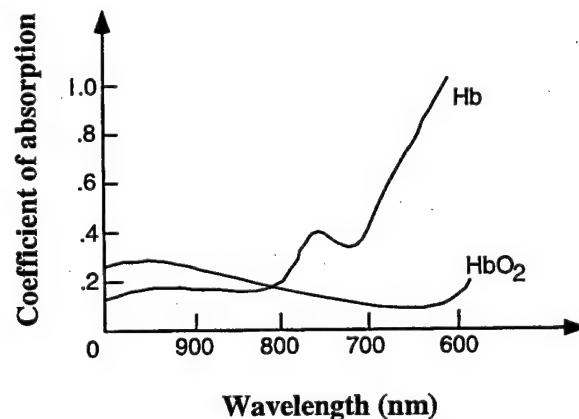

Figure 4: Effects of LED Wavelength

Figure 5 shows the original prototype touch sensor was fabricated by embedding a single phototransistor and red LED within a prefabricated plastic fingernail. The plastic fingernails

7

were then attached to the fingernails of the human using a thin strip of adhesive gum such as sticky-tack around the perimeter of the nail.
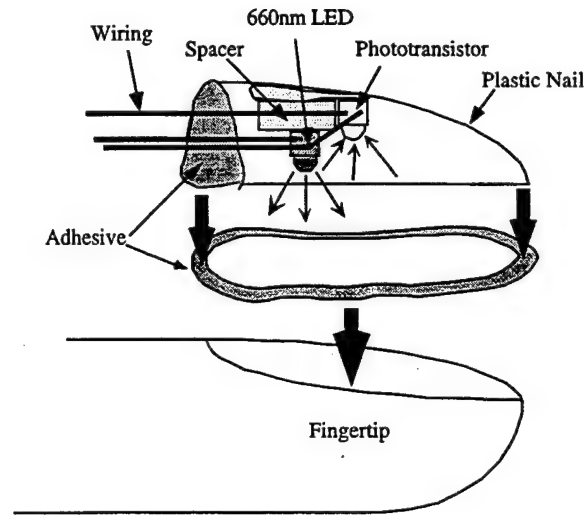

Figure 5: Prototype Design for the Fingernail Touch Sensor

3. Modeling and Analysis

3.1 Problem Formulation

Figure 1 showed the typical behavior of the color change of the fingernail with increasing touch force at the fingertip. In order to understand the correlation between the touch force and the observed behavior, we intend to create and verify a hemodynamic model of the blood flow in the fingertip. The goal is a model that can predict the magnitude and location of an input force, based on the measurable color-change output. Specifically, we wish to filter out color changes that are caused by bending of the finger. The basic behavior that must be portrayed by the model is as follows:

1. A mechanical force at the fingertips constricts the veinous return of blood
2. Constriction of the veins causes arterial blood to pool up under the fingernail
3. High force causes constriction of the arterial supply at the tip of the finger
4. Local constriction of the arterial supply causes blood to drain from the nail bed, resulting in a characteristic white band

8

At this stage we will make several assumptions:

1. The soft tissue of the fingertip can be modeled as a material with linear elasticity and damping.

2. The veins have a linear elasticity up to some sharp cutoff

3. Constriction of the veins modulates their fluidic resistance, affecting the pressure in the arterial supply under the nail

4. The arterial supply under the nail can also be modeled with linear elasticity and damping, but with discrete limit

5. The bone effectively shields the arterial supply under the rear portion of the nail from the direct influence of touch forces

6. The increase in blood volume under the nail effectively increases the path length over which light is absorbed by the blood

7. The effects of shear forces are neglected

8. Effects of oxygen concentration are neglected

3.2   Hemodynamic Modeling of Finger Tissue

As a first step towards a model that meets all of the above specifications, we will first create a lumped parameter model that explains the first two behavioral characteristics mentioned in the previous section. Figure 6 shows such a model. The arterial supply and veinous return are each modeled as a single channel of fluid flow through the finger, and are in series with each other. The walls of the arterial supply and veinous return are modeled as mass-spring-damper systems that expand and contract based on external forces and the local blood pressure. As they change diameter, their blood volume capacity and fluidic resistance change. A force on the soft tissue below the finger will constrict the veinous return. This will increase the fluidic resistance through the veins, driving the pressure to increase on the arterial side. The increase in pressure will cause the arterial supply to expand with blood, causing an effective increase in path length over which light is absorbed.
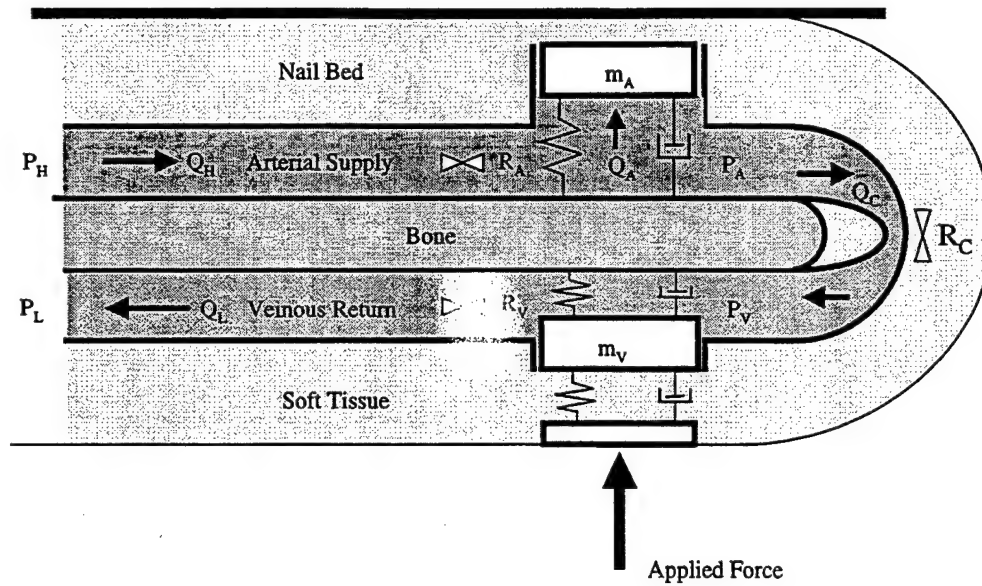
9

Figure 6: Hemodynamic Model of Fingertip

Dynamic state equations are written for this lumped parameter model, and simulations are performed to verify the required behavior.

4. Implementation

4.1 Prototype Design

Since the key to filtering out color changes induced by finger bending is to measure the pattern of the color change, the new prototype sensor is designed to use an spatial array of photodetectors along the longitudinal axis of the fingernail. Furthermore, by using three different wavelengths of LEDs, we can measure the relative concentration of oxygen in the blood as well as the change in blood volume (although the oxygen concentration is ignored in our initial model, we may wish to factor this in at a later time).

In order to fit multiple LEDs and photodetecting elements on the human fingernail in a non-obtrusive manner, we will take advantage of modern miniaturization technology. Figure 7 shows the miniature die form optical components that will be used for the sensor prototype. The photodetector arrays have 16 elements each, are 4x1mm in dimension, and have a broad spectral

range across the visible and infrared. The LEDs are 0.25x0.25mm and come in a variety of wavelengths. Our prototype will use a pair each of wavelengths 660nm, 770nm, and 940nm.

Figure 7: Miniature Optical Components for Sensor
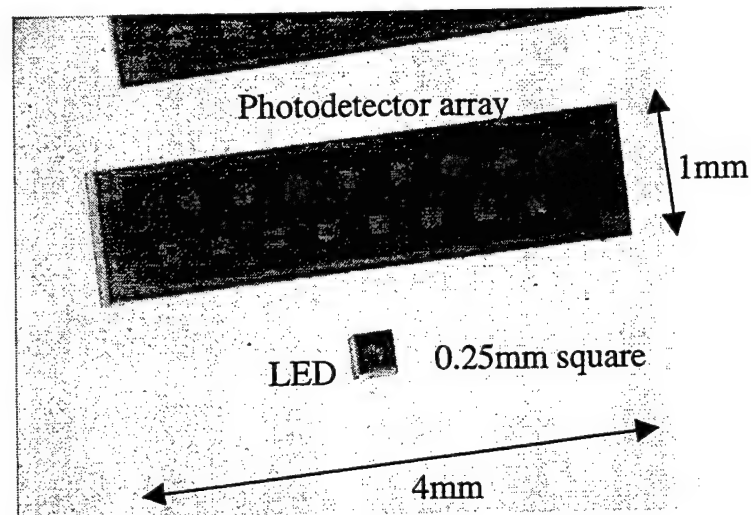
FINGERNAIL SENSOR FLEX CIRCUIT
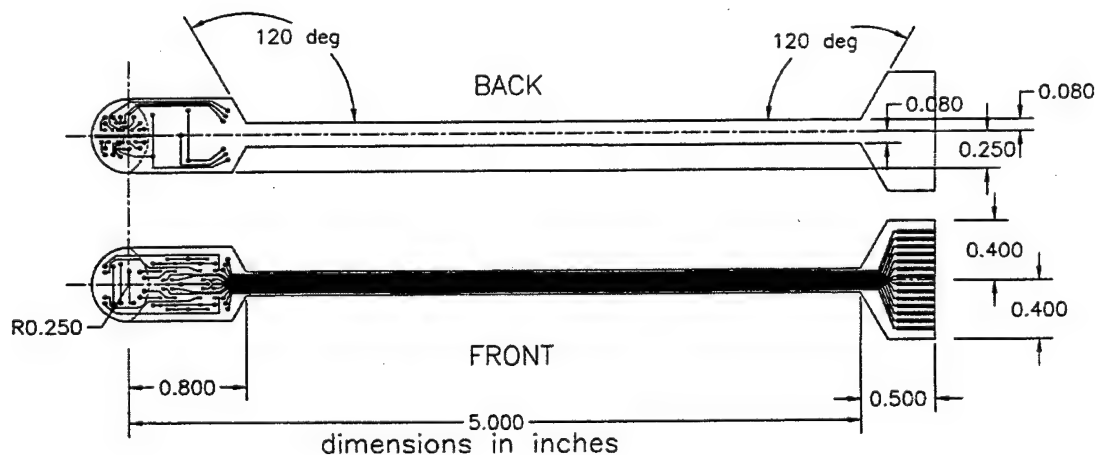
Figure 8: Design for New Fingernail Sensor

In order to adjust to the variability in fingernail contours, the optical components will be mounted on a flexible Kapton substrate, which can be bent to the shape of the fingernail before stiffening with epoxy. Figure 8 shows the layout for the flex circuit. The knobby portion on the

11

end will fit over the fingernail, while a thin flexible strip will reach back to the wrist, where a cable can be attached.

Figure 9 shows the layout of the optical components on the back of the flex circuit. Two of the photodetector arrays will cover the length of the fingernail. The LEDs are arranged symmetrically on either side at the center of the nail. The chips are attached with conducting epoxy and gold wire bonding.

Figure 9: LED and Photodetector Layout
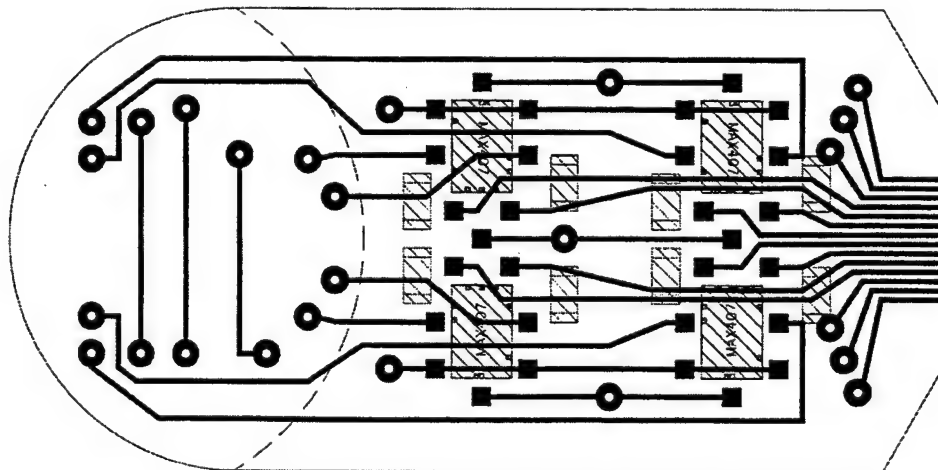
Figure 10: Circuitry Layout

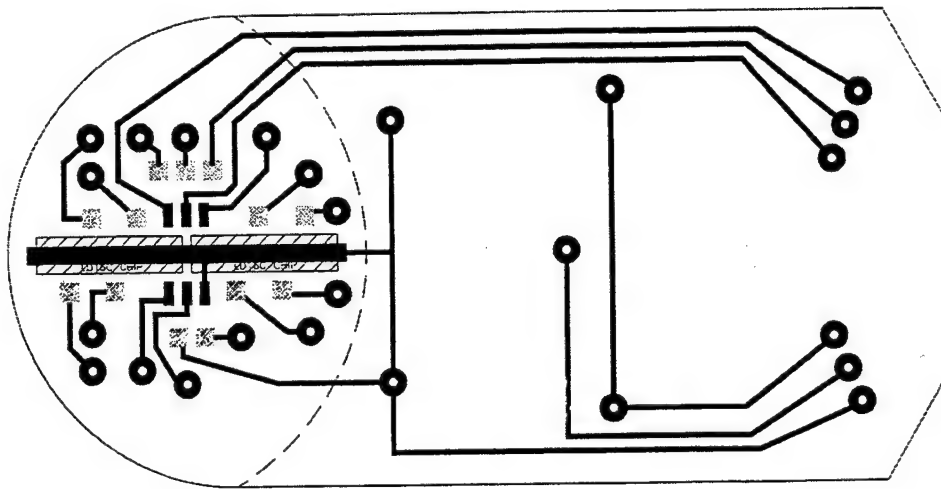Finally, Figure 10 shows the layout of the amplification electronics located on the front side of the flex circuit, which uses miniature op amp and resistor chips. The chips are attached using non-conducting epoxy and gold wire-bonding.

4.2 Experiments

In order to test the new nail sensor prototypes, a special apparatus was designed to allow the fingernail to be positioned at any angle with respect to an oscillating surface. A thin Archimedian force sensor on the surface measures the contact force as oscillatory displacements are imposed on the fingertip. The resulting output of the fingernail sensor can then be measured as a function of input in order to perform system identification.

Figure 11: Testing Apparatus for Nail Sensor

Figure 12 shows a typical photodetector output at the rear of the nail as a function of touch force applied normally to the bottom surface of the fingertip. More testing needs to be performed using output from the entire array of photodetectors for different wavelengths of light.

Figure 12: Experimental Data

5. Conclusions

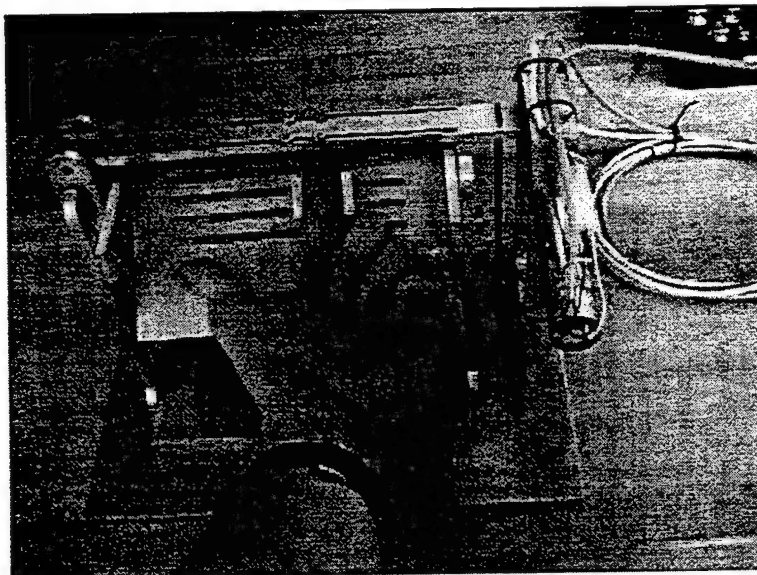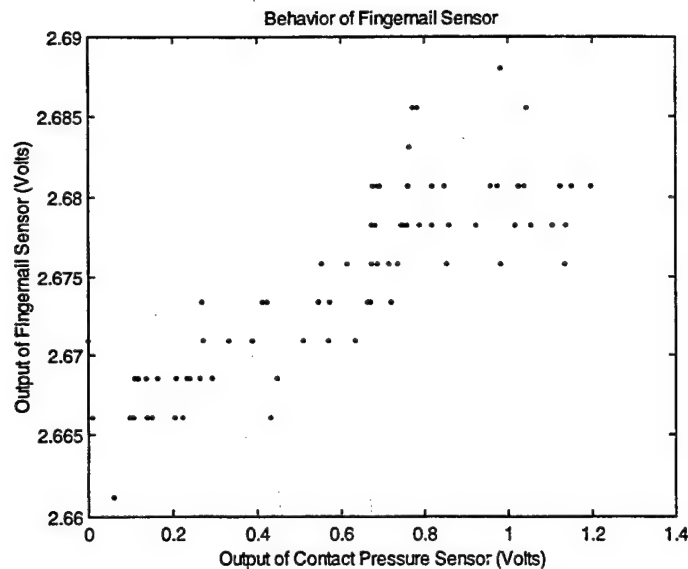A new type of touch sensor for detecting contact pressure at human fingertips has been presented. Fingernails are instrumented with arrays of miniature LEDs and photodetectors in order to measure changes in the nail color pattern when the fingers are pressed against a surface. Observable behavior of the color change phenomena was described and used to create a lumped parameter hemodynamic model, in order to understand the relation between touch force input and measurable color-change output. A new miniaturized prototype was designed, built, and tested.

Unlike traditional electronic gloves, in which sensor pads are placed between the fingers and the environment surface, this new sensor allows the fingers to directly contact the environment without obstructing the human's natural haptic senses. Several applications for this new technology are being investigated. These sensors could be used for the recording of chiropractic or surgical skills, where neither the patient nor the physician's fingers can be covered with sensors. In another paper [17], we propose the idea of a "virtual switch," whereby a traditional switch is replaced with merely an image of a switch on a surface. Measurements of a person's hand position, along with measurements of touch force from the fingernail sensor are used to activate the virtual switch. Because the virtual switches are merely images, they can be

14

located on all kinds of surfaces, such as the joints of a robot or on the surface of a desk, without modifying any hardware. They can be completely reconfigured by software and can have different functionality depending on which finger is used to press.

Appendix: Patent Search

A United States patent search was performed using Internet search engines in order to determine if any ideas similar to the fingernail touch sensor have already been patented. Our efforts did not turn up any existing patents for measurement of touch force using the fingernails. However, two patents related to fingernail sensing in general were found and may be of some use for increasing our own knowledge.

The first patent is for a "Method and apparatus for the automated identification of individuals by the nail beds of their fingernails" [18]. According to the patent, the nail bed of each person has a unique pattern of capillary loops, just like a fingerprint. By reflecting light off of the fingernail through arrays of optical fibers and measuring with CCD sensors, the capillary fingerprint of the nail bed can be determined and used to identify an individual. Their sensors use two different wavelengths of light. Although similar in functionality, their sensors are not worn on the human, but rather are free standing scanning devices. In comparison, the miniaturized wearable components we propose are significantly different from their design, and our purpose for measurement is also completely different.

The second patent is for a "Sensor for performing medical measurements, particularly pulsoximetry measurements on the human finger" using a sensor which is adhered to the fingernail [19]. The sensor is also patented in Japan, Germany, and EPO. The sensor is wearable and is adhered to the nail similarly to ours, but does not have an array of detectors and is only used for pulsoximetry. In comparison, our sensor is not concerned with the measuring the AC signal used for pulsoximetry, but rather the DC signal that results from touch force.

A thorough search was performed for any other patents having to do with some form of measurement on the fingernails, but none were found.

References

[1] D.J. Sturman and D. Zelzer, "A Survey of Glove-base Input", IEEE Computer Graphics & Applications, January, pp. 30-39, 1994.

[2] T.B. Sheridan, Telerobotics, Automation, and Human Supervisory Control, Cambridge, MA: MIT Press, 1992.

[3] R.S. Kalawsky, The Science of Virtual Reality and Virtual Environments, Addison-Wesley, 1993.

[4] G.C. Burdea, Force and Touch Feedback for Virtual Reality, Wiley, 1996.

[5] J. Kramer, and L. Leifer, "The Talking Glove: An Expressive and Receptive 'Verbal' Communication Aid for the Deaf, Deaf-Blind, and Non-vocal," tech report, Stanford University, Dept. of Electrical Engineering, 1989.

[6] D.J. Beebe, D.D. Denton, R.G. Radwin, and J.G. Webster, "A Silicon-Based Tactile Sensor for Finger-Mounted Applications", IEEE Transactions on Biomedical Engineering 45:2, pp. 151-159, 1988.

[7] J.G. Webster, Ed., Tactile Sensors for Robotics and Medicine, New York: Wiley, 1988.

[8] T.R. Jenson, R.G. Radwin, and J.G. Webster, "A Conductive Polymer Sensor for Measuring External Finger Forces", Journal of Biomechanics 24:9, pp. 851-858, 1991.

[9] J.S. Son, A. Monteverde, and R.D. Howe, "A Tactile Sensor for Localizing Transient Events in Manipulation", Proceedings of the IEEE International Conference on Robotics and Automation, pp. 471-476, 1994.

[10] S. Mascaro, K.W. Chang, and H.H. Asada, "Finger Touch Sensors using Instrumented Nails and their Application to Human-Robot Interactive Control", ASME IMECE, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, November, 1998.

[11] M. Brooks, "The DataGlove as a Man-Machine Interface for Robotics," 2nd IARP Workshop on Medical and Healthcare Robotics, Newcastle upon Tyne, UK, Sept., pp. 213-225, 1989.

[12] S.B. Kang, and K. Ikeuchi, "Grasp Recognition and Manipulative Motion Characterization from Human Hand Motion Sequences", IEEE Int. Conf. on Robotics and Automation 2, pp.1759-1764, 1994.

[13] C.P. Tung, and A.C. Kak, "Automatic Learning of Assembly Tasks using a DataGlove System", IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 1, pp. 1-8, 1995

[14] R.M. Voyles, and P.K. Khosla, "Tactile Gestures for Human/Robot Interaction," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 3, pp. 7-13, 1995.

[15] Gray's Anatomy: The Anatomical Basis of Medicine and Surgery, New York: Churchill Livingstone, 1995.

[16] R.V. Krstic, Human Microscopic Anatomy: an Atlas for Students of Medicine and Biology, New York: Springer-Verlag, 1991.

[17] S.Mascaro, K.W. Chang, and H.H. Asada, "Instrumented Fingernails: a Haptically Unobstructive Method for Touch Force Input," SPIE International Symposium on Intelligent Systems and Advanced Manufacturing, November 1998.

[18] A. Topping, V. Kuperschmidt, A. Gormley, "Method and Apparatus for the Automated Identification of Individuals by the Nail Beds of their Fingernails", US Patent 5751835, Issued May 12, 1998.

[19] S. Kaestle, M. Guenther, "Sensor for Performing Medical Measurements, Particularly Pulsoximetry Measurements on the Human Finger", US Patent 5776059, Issued July 7, 1998

Human-Machine Interface

CHAPTER 16

Human-Machine Interface and Interactive Control
Part 2: Virtual Switch Human-Machine Interface Using Fingernail Touch Sensors
    H. Asada, S. Mascaro

Human-Machine Interface and Interactive Control

Part 2: Virtual Switch Human-Machine Interface Using Fingernail Touch Sensors

H. Harry Asada
Professor, Principal Investigator

Stephen Mascaro
Graduate Research Assistant

Abstract

A novel human-machine interface using wearable sensors is presented. Fingernail sensors that detect touch forces at the fingertip are used as a means to acquire human intentions of pressing buttons and switches. Combined with a magnetic tracker detecting the position of the human hand, the nail sensor system can identify which switch the human wishes to push and when the switch has been pushed. This allows us to replace traditional physical switches, embedded in a wall, control panel, etc., by "virtual switches" that contain no electric circuit but are merely pictures showing the location of the switch. In the virtual switch system, the location and functionality of switches are determined by software and thereby changed flexibly depending on the progress of task performance, environmental conditions, and context. The virtual switch method is combined with a "hyper manual" storing task-procedure and operational information on a computer. Monitoring the human task performance allows the digital hyper manual to better guide the human, detect errors, and provide a safer environment.

1. Introduction

Electronic gloves have been extensively studied in the past decade in the robotics and virtual reality communities [1]. There are many ways of providing force feedback to the human from a virtual environment or from sensors on the robot, and many electronic gloves now make use of such force feedback [2][3][4]. A good example is the CyberGlove developed by Kramer [5]. However, few electronic gloves collect touch-force data from the human fingers as the human interacts with the environment. The ones that do make use of pressure sensing pads consisting of conductive rubber, capacitive sensors, optical detectors, or other devices which are placed between the fingers and the environment surface [6][7][8][9]. These sensor pads, however, inevitably deteriorate the human haptic sense, since the fingers cannot directly touch the environment surface. Recently, a new type of touch force sensor has been developed, which works by measuring the color change of the fingernail caused by touch forces [10]. This new sensor does not obstruct the natural haptic sense of the human. The introduction of such a sensor provides a new motivation for expanding the horizons of human-machine interaction.

Glove based input has been used increasingly in the last decade for teleoperation and other forms of human-machine interaction. Sturman and Zeltzer provide a comprehensive review of glove-based input [1], including applications to teleoperation and robotic control. Postural gesture recognition has been applied to teleoperation of robots and to teaching of robots by demonstration and guiding [11][12][13]. Recently, Voyles and Khosla developed a system for teaching-by-guiding by inferring human intentions from tactile gestures, which were measured by force sensors on the robot [14]. However, such as system is not very flexible, as it requires modification of the hardware on the robot. Instead, by measuring touch forces on the human side, we can achieve a much greater flexibility in measurement.

First in this paper, the principle of this virtual switch system is described and its features and utility are discussed. Practical embodiment of the concept by using fingernail sensors and magnetic tracker is presented. The method is applied to a human-robot interface for task programming and manual control. Effectiveness of the virtual switch method is quantitatively evaluated through experiments using human subjects. At last, the virtual switch method is

2

combined with a "hyper manual" storing task-procedure and operational information on a computer. Monitoring the human task performance allows the digital hyper manual to better guide the human, detect errors, and provide a safer environment. Application to the operation of a hybrid bed/wheelchair system for bedridden patients demonstrates the features of the combined virtual switch and hyper manual system.

2. Concept of Virtual Switches

2.1 Physical Switch vs. Virtual Switch

In the home as well as work environment, humans constantly supervise, control, and communicate with devices, computers and machines using a multitude of switches. Switches are rudimentary means for the human to communicate his/her intention to machines. The wearable finger touch sensor would replace the traditional switches and enhance the human-machine interface. Figure 1 depicts the functionality of a traditional switch and shows how the wearable finger touch sensor provides the same functionality and replaces the physical switch. As shown in Figure 1(a), the traditional switch works by means of:

1. The human movement of his/her finger to physically push some button of the switch
2. The detection of the human intention by electrical contact in the switch, and
3. Transfer of the detected signal to a specific part of the machine to change its state
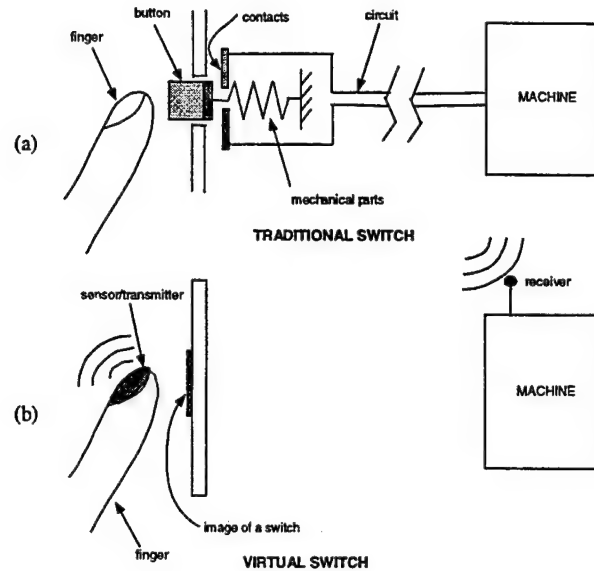
Figure 1: Traditional Switch vs. Virtual Switch

In the traditional switch, the detection of human intention is performed by a device that is physically attached and wired to a specific part of the machine. Traditional switches have the following limitations:

1. They generally cannot be reconfigured without reinstalling
2. They can be damaged in hazardous environments – chemical, mechanical, etc.
3. They take up space that could be used for some other purpose
4. They can be activated accidentally if bumped with something other than a finger

With the new wearable fingernail touch sensor, the ability now exists to measure human intention through touch without affecting a change on the environment. The detection of human intention can be performed by the device worn by the human rather than the one attached to the machine. To this end, we propose the concept of a "virtual switch." As shown in Figure 1(b), the virtual switch is not a mechanism, but is an image on a surface that represents a switch. The intention of activating the switch is detected by the fingernail touch sensor, coupled with 3-D finger position measurements from an electronic glove. The signal is transmitted wirelessly from the human to the machine. When a touch is detected on a finger whose position measurement corresponds to a certain virtual switch, that switch is activated. This will eliminate many of the

4

problems listed above. To summarize, the virtual switch panel offers the following advantages over traditional switches:

1. Virtual switches cannot be activated accidentally if bumped with something other than finger
2. Virtual switches can be rearranged and reconfigured without reinstalling
3. Virtual switches can have different functions for different fingers
4. Virtual switches can share the workspace and do not monopolize a work surface
5. Virtual switches can move around in space and change in functionality as a task progresses

The concept of the virtual switch panel opens up numerous possibilities for human-machine communication, and can be anything from a simple virtual on-off button to an entire virtual computer keyboard. Virtual switches can be placed at diverse surfaces such as walls, control panels and remote switch boxes, furniture, and even the body of the machine itself.

2.2 Human-Machine Interface

Figure 2 shows a sketch of one embodiment of the virtual switch panel. In this scenario, the human is working alongside a robot to accomplish a task. The human is wearing some form of electronic glove with open fingertips, which tracks the position of his fingers in 3-D space, and his fingernails are instrumented with the photo-reflective sensors to measure finger touch force. Virtual switches are painted on the surfaces around his workspace as well as the surface of his workspace. Perhaps some switches are even painted on the robot or human himself. Whenever, the fingernail sensor detects a sudden touch force, it relays the signal to the computer or robot controller along with the position of the finger which committed the touch. If the computer recognizes that the position corresponds to a certain virtual switch, then that switch is declared "activated." The function associated with the switch is performed, and the computer provides feedback to the human audibly or otherwise to confirm the activation of the switch. In this way the human can activate the robot, the computer, or other devices in his work area without affecting any change on the environment. Furthermore, the functions of each of the switches can be reprogrammed by the human at any time without having to do any work

5

mechanically. The virtual switches can even take on different functions automatically at different stages of a task, or have different functions depending on which finger activates them. Like a computer mouse with two buttons, different actions can be recognized by using multiple finger touch sensors. Finally, the human can work over top of the virtual switches and use his desk for other tasks without the switches getting in the way.
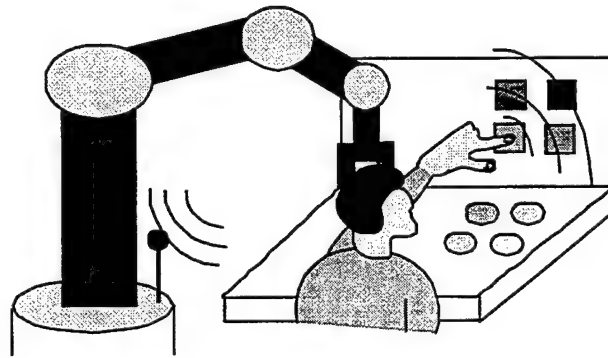
Figure 2: Virtual Switch Panel

Figure 3 shows the next level of this embodiment, which is the "totally virtual switch panel." This embodiment has all the features of the original virtual switch panel, only in this case the switches are not even painted or drawn on the surfaces of the workspace. Instead, the switches are either projected onto the workspace, or the human wears a head-mounted, heads-up-display, which superimposes computer images of the switches on his view of the workspace. By tracking head motion, the images can be made to appear stationary on a particular surface or move around in a desired fashion. Looking in different places can cause different switch panels to be displayed. Switches can be rearranged and reconfigured completely by software.

Figure 3: Totally Virtual Switch Panel

3. Manual Robot Control and Programming using a Virtual Switch Panel

3.1 System Construction

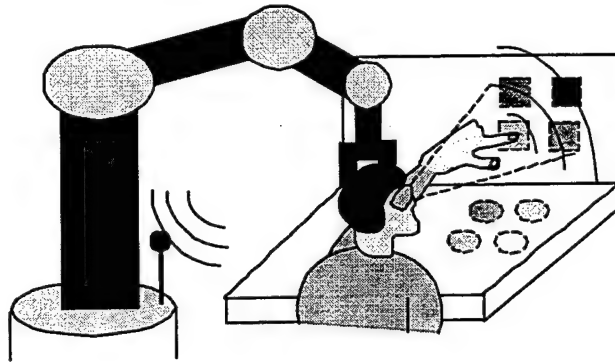Since the virtual switch has increased flexibility in terms of positioning and functionality, it seems reasonable to expect that increased performance can be achieved by making the virtual switch panel more intuitive than a traditional switch panel. A good case study for this hypothesis is the control of an industrial robot such as the PanaRobo KS-V20 manufactured by Panasonic. Figure 4 shows a picture of the robot and its teaching pendant. The robot has five rotational joints in series and can be controlled in joint angle mode, or linear Cartesian mode. As seen in Figure 4, the top three pairs of buttons on the right are used for linear Cartesian moves, and the top five pairs are used for joint angle moves.

A major limitation of this standard method of control is the lack of intuition between the buttons on the pendant and the resultant actuation of the robot. Depending on which way the person is facing the robot, a right button for example could cause the robot to move to the left. A button for a particular joint has no intuitive connection to the particular joint of the robot. There is much room for improvement in intuition by using a virtual switch panel. In particular, by using virtual switches, we now have the option of mounting switches on the workspace and the robot itself. Because the switches are virtual, they do not risk being damaged within the workspace environment, and they do not require any physical modifications to the robot hardware or the workspace.
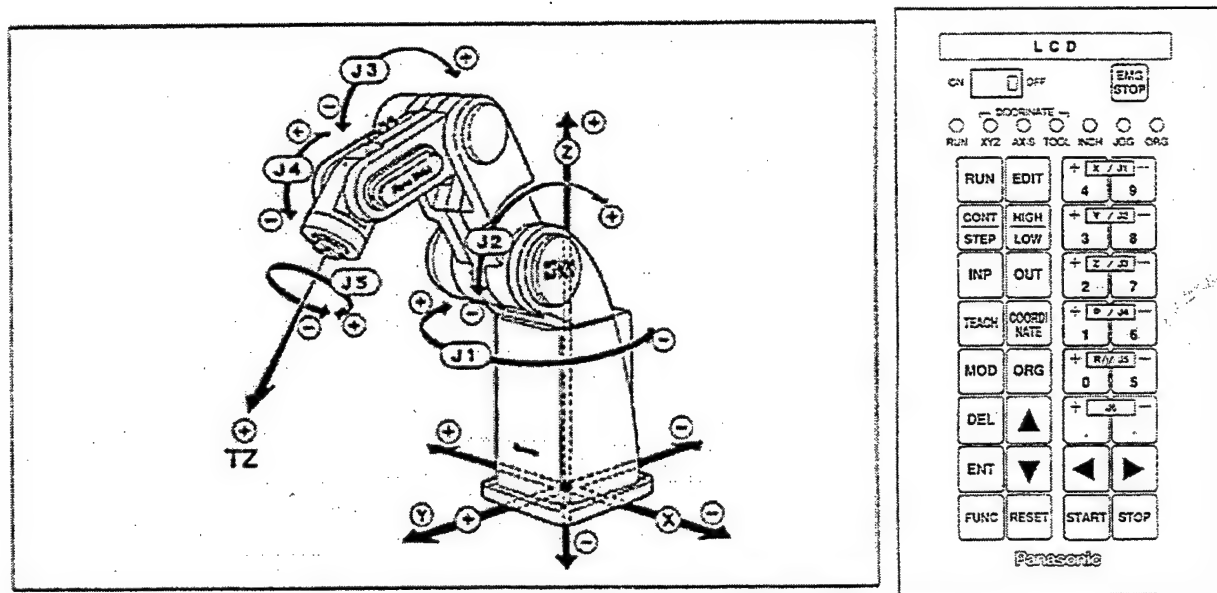
Figure 4: PanaRobo KS-V20 and Teaching Pendant
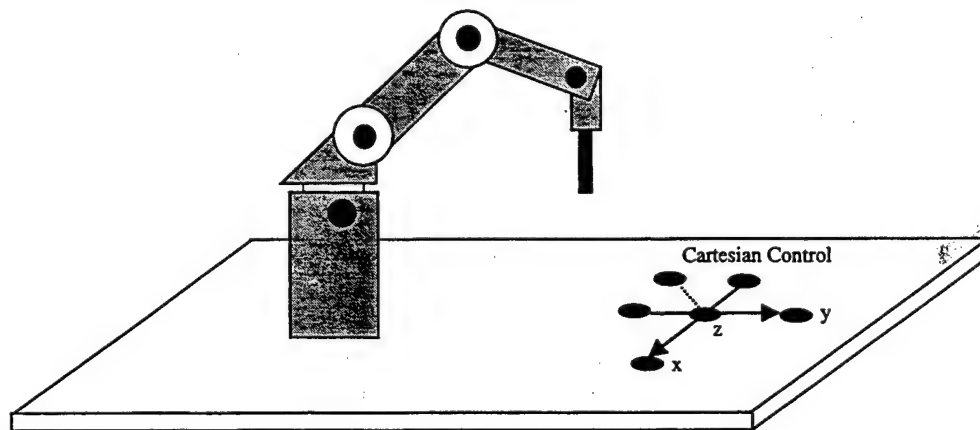
Figure 5: Virtual Switch Configuration for the Robot

Figure 5 shows an obvious configuration for the virtual switch panel for the KS-V20 robot. Virtual switches are located on each joint in order to perform joint moves, intuitively connecting each switch with its resulting actuation. The human is equipped with two touch force sensors on the index and middle fingers. The index finger causes counter-clockwise rotations while the middle finger causes clockwise rotations. Virtual switches for linear Cartesian control are located on the surface of the workspace in the shape of a set of coordinate axes,

8

corresponding to the coordinate frame of the robot. The operator can use two switches for each axis, or one switch for each axis with separate fingers for positive and negative directions.

We now present specific experiments to compare human performance using the virtual switch panel with performance using the traditional teaching pendant.

3.2   Experimental Evaluation

For this paper, we will focus on testing human performance using linear Cartesian control. An experiment was set up where a human operator is required to move the tip of the robot tool along a path in 3-dimensional Cartesian space. Figure 6 shows a drawing of the experimental setup. The operator is asked to maintain a constant gap of approximately 5-7 mm between the tip of the tool and the path surface. The total path length is 1.42 m. The path is comprised of 27 distinct moves: 11 in the x-direction, 9 in the y-direction, and 7 in the z-direction. The operator must therefore make at least 27 decisions in order to choose the correct switches, either on the teaching pendant or on the virtual switch panel. The robot is programmed to run at a linear speed of 25 mm/sec for both cases. The time it takes the operator to complete the course will be used as a measure of human performance for each method of control.
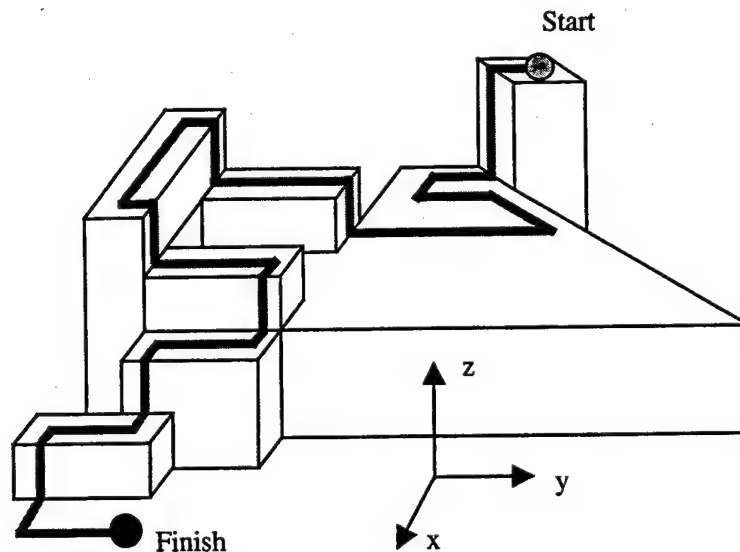
Figure 6: Tool Course

Figure 7 shows the results of this experiment for four human operators. As one might expect, there is certain variation between the operators. However, the learning curve trends are quite consistent. Compared to the teaching pendant, there is much less of a learning curve when using the virtual switches. The operators converge faster on their peak performance when using the virtual switches. This suggests that the virtual switches are a better match for the learning pattern of a human. Also, for some of the operators, the steady state performance is better when using the virtual switches as compared to the teaching pendant. This suggests non-intuitiveness cannot always be made up for by repetitive experience. The virtual switches can allow some people to operate at a consistently higher performance in tasks such as these.
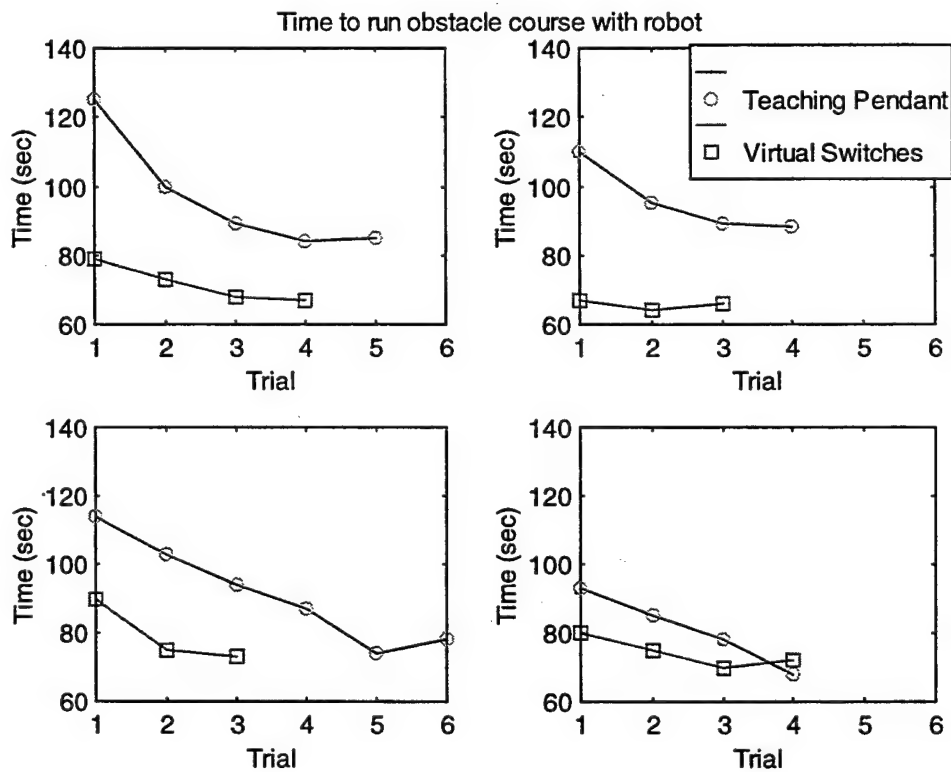
Figure 7: Performance Data for Teaching Pendant and Virtual Switches
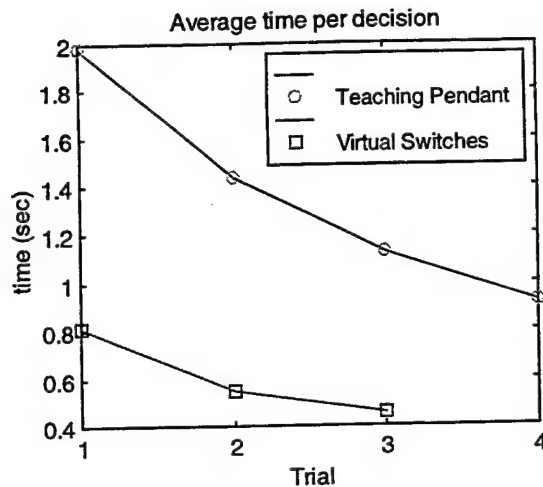
Figure 8: Composite Average Time Per Decision for Teaching Pendant and Virtual Switches

Based on the speed of the robot, the minimum time to cover the distance of the course is 57 seconds. By subtracting this minimum time required for moving, we can get the approximate time taken up by the mental decision making for each operator. Dividing by the total number of decisions required, we get the approximate average time per decision for the operators. The results are averaged for the four operators and shown in Figure 8. The difference here is quite pronounced. The virtual switch seems to offer an initial advantage of less than half the time per decision compared to the teaching pendant. At most, about half the number of trials is required for the learning curve to level off. To make any general comparisons about the final steady-state performance, this experiment needs to be conducted with larger number of trials on larger pool of human operators.

4. Digital Manuals with the Virtual Switch Human-Machine Interface

4.1 Integration of Manual Information with Human Behavior Understanding

The flexibility that the virtual switch system provides would allow us to assist the human in performing complex tasks that need interactive control and coordination between a machine and the human. In this section we explore the possibility of using virtual switches for performing

a complex procedure described in a manual. Different portions of virtual switches are to be activated and presented to the human following the procedure declared in the manual.

In general, a manual describes step-by-step instructions for operating a machine to perform a certain task. Typically a human follows a procedure described in a manual, which includes a sequence of operations, usually pressing buttons and knobs. In the past decade, many manuals have been computerized, i.e. digital manuals, for better service and easier use. The functionality, and features of digital manuals can be furthered by combining wearable sensors such as the finger touch sensors and hand position sensors, which monitor human behavior. Namely, combining virtual switches with a digital manual would create a powerful aid for guiding a user through a complex operational procedure. Figure 9 illustrates such a combined system, called a "wearable hyper manual".
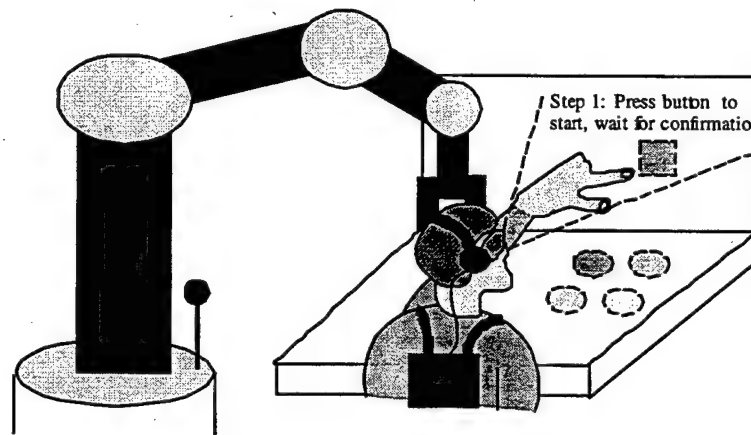
Figure 9: Wearable Hyper Manual

The wearable hyper manual consists of a wearable computer storing manual information, wearable and/or stationary sensors monitoring human behavior, and a display and/or headset for providing instructions to the human. Unlike traditional digital manuals, where the user must retrieve items of information needed for each step of operation, the wearable digital manual system monitors the human behavior, identifies in which stage of procedure the human is currently involved, and provides the right items of instruction needed for that stage. Furthermore, inputs from the human are acquired from the virtual switch panel described above.

The virtual switch panel presented to the human would be varied depending on the stage of the procedure and the relevance to the context of the task. The execution of this entire process entails a task programming and process control engine. Such an engine would represent the task, code the procedure, recognize each task stage, observe human behavior, retrieve manual information, present instructions, display control panels, and acquire human inputs to coordinate a target machine with the human inputs.

4.2 Application to a Hybrid Bed/Wheelchair System

To begin with, consider the virtual switch system applied to a hybrid bed/wheelchair system for bedridden patients, as shown in Figure 10. This hybrid bed/wheelchair, called RHOMBUS (Reconfigurable holonomic omnidirectional mobile bed with unified seating), was developed to eliminate transfer between a bed and wheelchair [15]. This requires a certain procedure for converting the bed to a chair and vice versa. To alleviate difficulties and assure safety, virtual switches are placed at various surfaces on the chair and bed. The system is to be operated by a caregiver wearing the finger touch sensors and hand position sensor. Numerous virtual switches can be imbedded in the bed/chair surface for acquiring the caregiver's intention. For example, when the caregiver wishes to raise the back leaf of the bed/chair system, he/she touches the back side of the back leaf and tends to push it upward. The virtual switch imbedded in the back leaf recognizes the human motion by detecting the hand location and the pressure increase at his/her fingers.
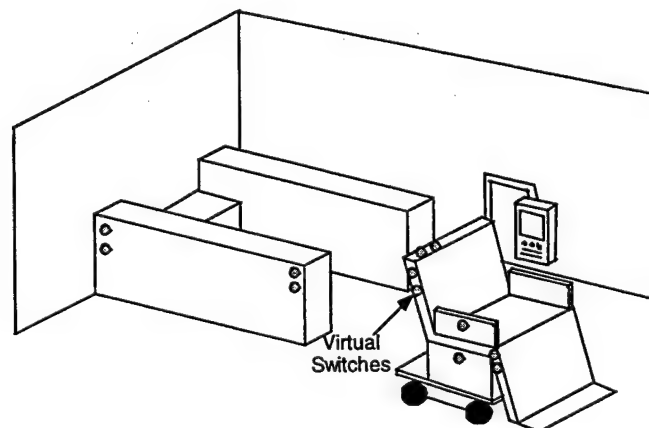

Figure 10: Distributed Virtual Switch System

The detected signal is then transmitted to the powered bed/chair for activating the actuator raising the back leaf. Let us suppose that, after raising the back leaf, the caregiver wants to push the bed (now reconfigured to a wheelchair) forward. The virtual switch imbedded in the wheelchair handle detects the caregiver's intention, when he/she touches the wheelchair handle and tends to push it forward. The position sensor recognizes that the caregiver places his/her hand on the handle, and the finger touch sensors detect that the forward button printed on the handle is pressed by his/her fingers.

The assignment of virtual switches to individual portions of the bed/chair system can be changed depending on the context, situation, and stage in the task. For example, virtual switches during the bed-mode operations and the chair-mode operations may be altered by simply changing the "map" relating sensor signals to the control actions. Although pressing the same point of the machine, different actions can be generated. Pressing the back leaf, for example, is recognized as the intention of changing the back leaf angle only when the operation is in the bed mode. Pressing the same back leaf during the wheelchair mode creates no action, thus avoiding erratic operations. This context-dependent assignment is extended to the concept of "wearable digital manuals".

5. Conclusions

In this paper we proposed the idea of a "virtual switch," whereby a traditional switch is replaced with merely an image of a switch on a suface. Measurements of a person's hand position, along with measurements of touch force from the fingernail sensor are used to activate the virtual switch. Virtual switches offer an increased flexibility in placement and functionality over traditional switches. The virtual switch concept was tested against a traditional teaching pendant for teleoperated teaching of a robot. Experiments show that the flexibility in designing virtual switches can result in an increased intuition for the human operator. Initial decision-making time as well as learning time can be reduced by over 50%.

The idea of a wearable hyper manual was also presented, which can be combined with the virtual switch method to prompt a human operator with instructions while interacting with the

machine. This entails a task programming and process control engine that would represent the task, code the procedure, recognize each task stage, observe human behavior, retrieve manual information, present instructions, display control panels, and acquire human inputs to coordinate actuation of the machine with the human inputs.

References

[1] D.J. Sturman and D. Zelzer, "A Survey of Glove-base Input", IEEE Computer Graphics & Applications, January, pp. 30-39, 1994.

[2] T.B. Sheridan, Telerobotics, Automation, and Human Supervisory Control, Cambridge, MA: MIT Press, 1992.

[3] R.S. Kalawsky, The Science of Virtual Reality and Virtual Environments, Addison-Wesley, 1993.

[4] G.C. Burdea, Force and Touch Feedback for Virtual Reality, Wiley, 1996.

[5] J. Kramer, and L. Leifer, "The Talking Glove: An Expressive and Receptive 'Verbal' Communication Aid for the Deaf, Deaf-Blind, and Non-vocal," tech report, Stanford University, Dept. of Electrical Engineering, 1989.

[6] D.J. Beebe, D.D. Denton, R.G. Radwin, and J.G. Webster, "A Silicon-Based Tactile Sensor for Finger-Mounted Applications", IEEE Transactions on Biomedical Engineering 45:2, pp. 151-159, 1988.

[7] J.G. Webster, Ed., Tactile Sensors for Robotics and Medicine, New York: Wiley, 1988.

[8] T.R. Jenson, R.G. Radwin, and J.G. Webster, "A Conductive Polymer Sensor for Measuring External Finger Forces", Journal of Biomechanics 24:9, pp. 851-858, 1991.

[9] J.S. Son, A. Monteverde, and R.D. Howe, "A Tactile Sensor for Localizing Transient Events in Manipulation", Proceedings of the IEEE International Conference on Robotics and Automation, pp. 471-476, 1994.

[10] S. Mascaro, K.W. Chang, and H.H. Asada, "Finger Touch Sensors using Instrumented Nails and their Application to Human-Robot Interactive Control", ASME IMECE, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, November, 1998.

[11] M. Brooks, "The DataGlove as a Man-Machine Interface for Robotics," 2nd IARP Workshop on Medical and Healthcare Robotics, Newcastle upon Tyne, UK, Sept., pp. 213-225, 1989.

[12] S.B. Kang, and K. Ikeuchi, "Grasp Recognition and Manipulative Motion Characterization from Human Hand Motion Sequences", IEEE Int. Conf. on Robotics and Automation 2, pp.1759-1764, 1994.

[13] C.P. Tung, and A.C. Kak, "Automatic Learning of Assembly Tasks using a DataGlove System", IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 1, pp. 1-8, 1995

[14] R.M. Voyles, and P.K. Khosla, "Tactile Gestures for Human/Robot Interaction," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 3, pp. 7-13, 1995.

[15] Mascaro, S., Spano, J., and Asada, H. "A Reconfigurable Holonomic Omnidirectional Mobile Bed with Unified Seating (RHOMBUS) for Bedridden Patients," IEEE Int. Conf. on Robotics and Automation, Albuquerque, New Mexico, April 1997.

Home Networking & Wireless Communication

CHAPTER 17

Home Automation Network - HANET
S. Martel, S. Lafontaine, I.W. Hunter

d'Arbeloff Laboratory for Information Systems and Technology
MIT

HOME AUTOMATION NETWORK - HANET
PROGRESS REPORT FOR YEAR 1

Sylvain Martel, Serge Lafontaine and Ian W. Hunter

This report describes the first year development of a new home automation and health care network designated HANET (Home Automation Network).

Introduction

Objectives

The home automation network under development will provide the interconnection structure necessary to fulfill the requirement of advanced home-integrated functionality and services, including home health care. The suitability of HANET for performing such tasks within the home environment has been described in the recent report: *Home Automation Network - HANET, Second Year Research Plan.*

HANET Protocol Architecture

The architecture of HANET relies on various layers namely, the physical layer, the link layer, the transaction layer, the network management layer, and the HANET protocol layer. All layers but the HANET protocol layer are based on the IEEE-1394 (FireWire) Standard. The physical layer although fully compliant with the IEEE-1394 Standard, uses a new architecture with special additional embedded functions that enhances the suitability of the system for the home environment. The HANET protocol specifications are completely independent of the IEEE-1394 Standard but are encapsulated within the payload of network packets that are fully compliant with the IEEE-1394 Standard.

The physical layer is required for devices and networking connectivity, transmission media, arbitration, data re-synchronization, network initialization, encode/decode functions, and signal levels. The link layer is used for packet transmission/reception and

1

cycle control. The transaction layer supports service primitives such as request, indication, response and confirmation. The network management layer is responsible for the network resources management and allocation including the isochronous resources (see IEEE Std. 1394-1995 document for more details). The HANET protocol layer contains all mechanisms necessary to transfer commands, data, and status information that are pertinent to the home automation and health care.

First Year Research and Development

Main Objective

The main objective of the first year research plan for HANET was primarily the development of the hardware framework that would allow us to validate the various ideas and concepts embedded in our proposed home automation and health care network. As such, four types of electronic boards or modules were developed. Short descriptions of these modules were provided in the first year research plan. These modules are:

1. FWM: Intelligent home outlet and hub running at 200 Mbps and used to interconnect the devices.
2. UCM: Digital signals processing board with special interfaces and IEEE-1394 link and transaction layers capabilities.
3. ADM: Analog-to-digital conversion module that links to the UCM.
4. DAM: Digital-to-analog module that links to the UCM.

All these modules have been built, primary tests have been performed, and a fair amount of embedded software has been developed. It is fair to say that all objectives for the first year have been met since we are confident at this stage that these modules are operationally ready for the next development phase which has been described in the recent report: *Home Automation Network - HANET, Second Year Research Plan.*

New Intelligent Home Outlet and Hub - FWM

The FWM is a first prototype used to validate the idea of intelligent outlets and hubs within the future home. An intelligent outlet as envisioned in this report is a unit which can be installed throughout the home to provide connectivity to HANET for a multitude of intelligent devices that comply with the IEEE-1394 Standard. The new Sony's cameras are just one example of such new IEEE-1394 compliant devices and the number and types of devices that comply with the standard are likely to increase rapidly in a very near future.

The present implementation of the intelligent outlet has been fully tested and 20 units have been built. The primary tests consisted of passing real-time (30 frames/s) video images recorded from a Sony camera to a personal computer through several FWM outlets. The recording images were displayed in real-time on the computer's screen. The hot swapping capability was also tested as well as all the ports on the FWM outlets. The speed and the quality of the images displayed on the screen were very good without any observable glitches. The tests were performed at a transmission rate of 200 Mb/s.

Each outlet has three IEEE-1394 ports and as such, it can be used as a hub making the implementation of various network topologies in the home much easier without the need for additional hardware. This capability of the FWM has been also fully tested.

The FWM can also act as either a network repeater and/or a physical layer for another module containing the link layer capability without a physical layer. The FWM has another special port that can connect directly to a link layer controller located in another module such as the UCM. The advantage of such a scheme is the elimination of an extra node to provide connectivity. This allows more interconnections through the same network since although relatively high, there is always a limitation on the number of nodes that can be connected on a network. More importantly, this scheme may prevent short network initialization time to take place when a new device is connected or disconnected. The circuit that allows detection of the presence of a link layer controller and the switching capability between the physical layer and the repeater modes as well as the connection port

between the link layer and the FWM physical layer have also been fully tested and are working.

In summary, the intelligent outlets and hubs with the functionality described above are fully operational.

Special Hardware for Experimentation

A network alone is not enough to provide all necessary framework to test and validate new home automation and heath care applications. For instance, a network only provides interconnection between nodes but does not interface to various sensors and/or actuators most often used in these applications. Instead of developing special hardware compatible with the intelligent outlet for each application, the strategy was to develop a flexible yet powerful system that would link to the FWM intelligent outlet. The new hardware would provide the processing power and digital/analog interfaces necessary to support and experiment a wide variety of applications.

With this approach, the time consuming and expensive process of developing and testing hardware and system's software for each application would be entirely eliminated or at least minimized. As such, we developed during the first year three electronic modules namely, the Universal Controller Module (UCM), the Analog-to-Digital conversion Module (ADM), and the Digital-to-Analog Module (DAM).

The UCM contains all the electronics required to implement the link and transaction layers as described previously. Both asynchronous and isochronous transactions can be supported. The UCM can link directly to one of the FWM intelligent outlet. The UCM also provide 40 MFLOPS (Millions Floating Points Operations per Second) and 48 digital I/O channels.

The ADM and the DAM provides high-resolution Analog-to-Digital (A/D) and Digital-to-Analog (D/A) conversions capabilities respectively. Both modules can be "piggy-back"

4

connected directly to the UCM. The three modules with the FWM forms a complete self-contained unit providing most of the hardware framework necessary to quickly implement and validate new home automation and health care applications.

All of these modules have been designed; several units have been built and programmed. Although more software needs to be implemented on the UCM, the codes to perform some basic functions have been developed and tested. Most of the software development is planned for the second year of the project (see the recent report: *Home Automation Network - HANET, Second Year Research Plan*).

Conclusion

The objectives of the first phase have been met with the development and accessibility of the hardware framework necessary to proceed to the second phase of the project, which has been described in the recent report: *Home Automation Network - HANET, Second Year Research Plan*).

Phase 2 Progress Report: October 1, 1998
Total Home Automation and Healthcare/Elder Care Consortium

Home Networking & Wireless Communication

CHAPTER 18

Minimum Energy Coding with Application to RF Transmission
H. Asada, K-Y Siu, C. Erin

Minimum Energy Coding with Application to RF Transmission

H. Harry Asada
Principal investigator

Kai-Yeung Siu
Co-investigator

Cem Erin
Graduate Research Assistant

Abstract

Energy efficient RF transmission is accomplished by devising a novel source coding algorithm for symbols with known statistics. Throughout the history of communications, the trend has been towards faster transmission. Thus, traditional source coding aims to achieve maximum compression to optimize transmission rate. Recent popularity of wireless communication devices, however, is requesting energy efficient RF transmission techniques. In this report, we first reformulate the source coding problem for energy efficient wireless transmission. Next, we propose a novel coding algorithm (ME Coding) that achieves optimal energy performance by taking advantage of the source symbol statistics. We prove the optimality of the code and compare the energy efficiency of ME Coding against traditional PCM techniques. We continue by identifying the parameters determining optimal performance and derive an optimality bound. Then, we introduce concatenation as a technique that further improves the optimal ME Coding performance through memory utilization. We elucidate the concatenation mechanism and present the achievable performance improvements. Finally, we propose a bound on optimal concatenation performance. We conclude that energy efficiency of a wireless communication system can be optimized through proper source coding.

Table of Contents

List of Figures

1 Introduction

1.1 Motivation

In communication theory, emphasis has predominantly been on bandwidth utilization and error correction, while energy efficiency has received little attention. Recent proliferation of wireless communication devices, however, is demanding new technology for energy efficient transmission [1, 2]. Since battery power is still limited and does not meet the demands for transmission over long distances and long time periods, energy efficient techniques must be developed for wireless transmission.

1.2 Background

Current energy efficient methods have evolved into three major research areas. The first area is concerned with network-level protocol design for reducing on-time of multiple devices [3]. The second area investigates architectural techniques at the device-level in order to optimize the on-time of various circuitry within the computation unit and the transmitter [4]. The third area concerns specifically with the computation unit. Within this area a group of researchers aim to maximize energy efficiency via sophisticated shutdown strategies and voltage scaling methods [5]. Another group recognizes that a major source of computational energy consumption is bit switchings. Hence, this group investigates various address coding and instruction scheduling techniques to minimize energy consumption by minimizing the number of bit switchings [6]. Recently, some work interconnecting the computation and transmitter energy consumption has appeared [7, 8]. These work investigate the tradeoffs between computational and transmission energy consumption.

1.3 Proposed Method

A major energy consuming component in a wireless device is the RF transmitter. Previous research on RF transmitter energy optimization includes high-level on-time reduction techniques, which are of limited advantage in continuous operation. In an attempt to optimize the transmitter energy efficiency in continuous operation, we propose a new approach and a new realm of research, that is, *design of source codes for energy efficient wireless transmission.* The goal is to find special source codes that minimize energy consumption, while satisfying the transmission rate requirements. We will develop a novel coding algorithm, minimum energy coding (ME coding), that minimizes transmitter energy consumption by taking advantage of the source symbol statistics. For optical transmission, an energy-saving code has been proposed in conjunction with a speech transmission system [9]. In this paper, we will develop the op-

timal energy coding algorithm for wireless transmission and prove the optimality of the code. Furthermore, we will determine all the parameters relevant to optimality and propose bounds on the optimal performance. For readers not familiar with source coding, an introduction to source coding is provided in Appendix A.

1.4 Application

This research aims to develop a novel communication protocol that will be applicable to all energy-critical wireless communication systems. As an immediate application, however, ME Coding is developed in conjunction with the finger ring sensor. The ring sensor continuously senses and transmits the physiologic data of multiple and remotely-located patients to a home computer that is connected to a nursing center or a central monitoring station [8].. The three major power consuming components involved in the ring sensor are *(i)* LEDs and photodiodes that acquire the patient's physiological information, *(ii)* the CPU that performs the computational operations, and *(iii)* the RF transmitter that emits RF signals representing the patient's data. We will concentrate on the RF transmitter consuming more than 40% of the total power consumption. A formulation of the RF transmitter power consumption will be given in the context of the wearable monitoring system, although it applies to a broad class of transmitters.

1.5 Overview of Report

The outline of this report is as follows. Section 2 introduces RF transmitters and formulates the power consumption problem in RF transmission. ME coding algorithm and the bounds of energy savings are discussed in Section 3. Section 4 introduces extensions of ME coding and derives the accompanying optimality bound. The report closes with some conclusions and recommendations in Section 5.

2 RF Transmitter Power Consumption

2.1 RF Transmitter

In a wireless communication system, the RF transmitter modulates the information to be communicated onto a carrier, amplifies the waveform to the desired power level, and delivers it to the transmitting antenna [10]. The transmitter includes a radio frequency oscillator that is modulated by the message signal, the bit stream of encoded data created by the CPU. The modulated signal is then multiplied in frequency up to the desired transmitting frequency and is amplified to the desired power level in the power amplifier.

The transmitter topology illustrated in Figure 1 is only one of many types. The modu-

V_{M} → Modulator → Frequency Multipliers → Power Amplifier → Antenna; Oscillator → Modulator
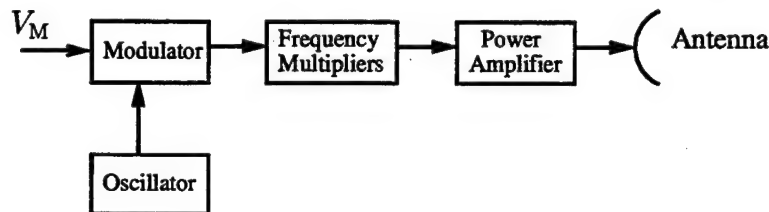
Figure 1: Block diagram of an RF transmitter

lation can actually take place in the power amplifier. Transmitter topology depends on the type of modulation used and the necessary power level. Narrow-band transmitters actually employ pulse, amplitude, or frequency modulation. Wide-band transmitters use single-sided or multi-mode modulation and are used for long-range military, marine, aircraft, and amateur communications. Many transmitter circuits are similar, all require low-noise amplifiers and oscillators.

The major power consuming transmitter component is the oscillator [10]. A harmonic oscillator is a circuit that outputs nearly sinusoidal signals for non-periodic input signals. Figure 2 shows the circuit diagram of the actual oscillator used in the finger ring sensor. This oscillator is a variation of the standard Colpitts oscillator. The Colpitts oscillator uses a bipolar junction transistor that is connected as a common-emitter amplifier with the base potential divider and the emitter resistor providing a DC bias [11]. The emitter resistor is decoupled to provide maximum gain. The collector load is a radio frequency choke. This is simply a large inductive impedance over the frequency range of interest. Positive feedback is provided by an LC circuit. Figure 2 shows that the ring sensor oscillator incorporates a quartz crystal. The crystal serves to rectify the resonant frequency of the oscillator, thereby increasing the accuracy of the frequency of oscillation. With the receipt of a high bit from the CPU, the LC circuit oscillates at its resonant frequency, which actuates the crystal. This results in an RF signal being emitted

6
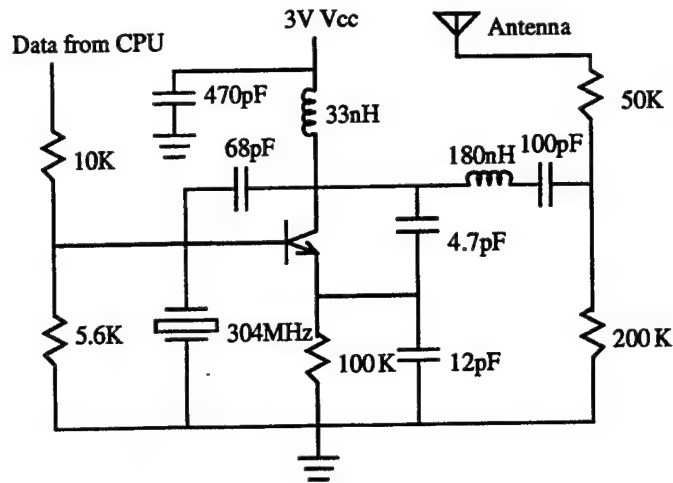
Figure 2: Colpitts Oscillator

from the antenna. The RF signal exhibits a short transient behavior followed by steady-state sinusoidal oscillation at 304 MHz. The resulting RF signal is shown in Figure 3.
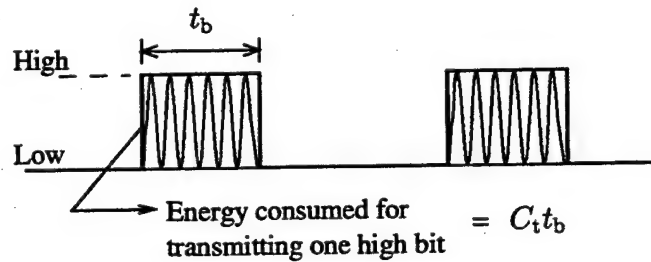
Figure 3: Digital RF Transmitter

2.2   RF Transmitter Power Consumption Formulation

Since the oscillator is actuated upon the receipt of a high bit only (see Figure 3), power consumption in the transmitter occurs only when high bits are sent and virtually no power is consumed when low bits are sent. The bit period is assigned a minimum detectable value that is determined from the channel characteristics. Since this bit period value ($t_b \approx 3.33 \times 10^{-3}$ sec) is much larger (a factor of nearly 10^6) than the oscillator period, the oscillator transients are neglected. Furthermore, the sinusoidal oscillator behavior results in the total RF transmitter power consumption, C_t, being constant. Consequently, the total power consumption for transmitting one high bit is proportional to the bit period, t_b. Hence, the average total power

consumption of the RF transmitter, C_{ave}, is given by

$$C_{\text{ave}} = (mC_t t_b)n_{\text{ave}}, \tag{1}$$

where m is the number of symbols transmitted each second, t_b is the bit period, and n_{ave} is the average number of high bits in each codeword. The average number of high bits in a codeword is given by

$$n_{\text{ave}} = \sum_{i=1}^{q} n_i P(s_i), \tag{2}$$

where n_i is the number of high bits in the ith codeword, $P(s_i)$ is the probability of the source symbol corresponding to the ith codeword, and q is the number of source symbols.

2.3   Power Consumption Optimization Strategies

Equation (1) suggests several strategies for maximizing energy efficiency (i.e., minimizing the power consumption). Energy efficiency can be improved by *(i)* decreasing the number of data points transmitted per second, m, *(ii)* optimizing the transmitter circuitry to minimize the total transmitter power consumption, C_t, and *(iii)* minimizing the bit period, t_b. Assuming the ring sensor is operating with these parameters having their optimal values, a further effective way of improving energy efficiency is decreasing the average number of high bits per codeword, n_{ave}. Average number of high bits per codeword, as defined in (2), depends on the number of high bits in each codeword, n_i, and the symbol probabilities, $P(s_i)$. If we have complete control over the mapping between the source symbols and available codewords, we can assign codewords having less high bits to symbols with higher probability. This strategy of assigning codewords to source symbols characterizes the ME coding algorithm. *Hence, ME coding can be considered to put the transmitter into a maximum sleep mode at the source encoding level.*

2.4   ME Coding Example

We will now demonstrate the ME coding algorithm using actual ring sensor data shown in Figure 4.

When the data is quantized such that the source alphabet is $S = \{0.25(j-1);\quad j = 1, ..., 18\}$, the symbol probabilities become those in Table 1 . We are also given that we want to transmit at least m symbols per second and that the physical medium allows a minimum bit period of t_b.

We begin by determining the codeword length, L, we will use. Since we want to encode 18 symbols, we should have $2^L \geq 18$. Furthermore, since we want to transmit at least m points
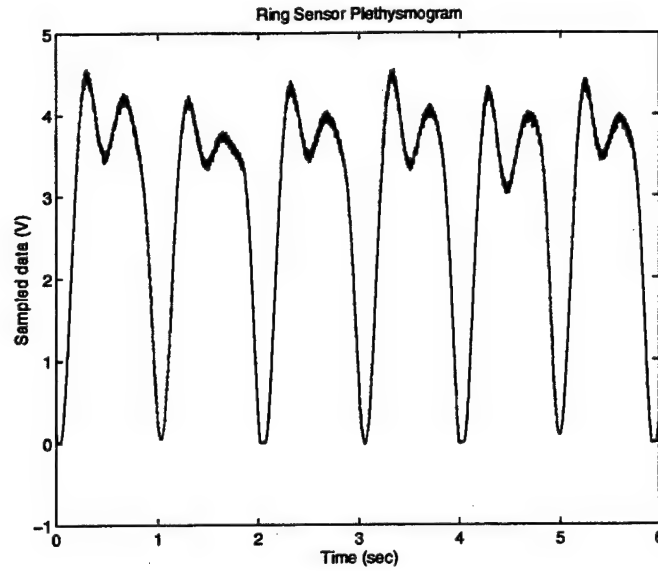
Figure 4: Ring sensor data

per second, we should have $L \leq 1/(mt_b)$. From these two constraints, we select the largest allowable L, so that we will have more codewords with less high bits. Let us assume that the resulting codeword length is $L = 5$.

Symbols	Prob.	ME	Symbols	Prob.	ME
2.75	0.191	00000	0.50	0.028	10001
3.00	0.185	00001	2.00	0.027	00110
2.50	0.158	00010	0.75	0.026	01010
3.25	0.068	00100	1.75	0.023	10010
3.50	0.062	01000	1.00	0.022	01100
2.25	0.053	10000	1.50	0.017	10100
3.75	0.045	00011	1.25	0.015	11000
0.25	0.031	00101	4.00	0.015	00111
0.00	0.029	01001	4.25	0.005	01011

Table 1: Ring sensor source symbols and their corresponding ME codewords

Now that we have selected the codeword length, we can start the ME coding algorithm. We begin by assigning the codeword of all low bits to the symbol with the highest probability (2.75). Then, we assign the $\binom{5}{1}$ available one-high-bit codewords to the next set of highly probable symbols. As shown in Table 1, the second through sixth symbols have codewords with only one high bit each. We continue by assigning the $\binom{5}{2}$ two-high-bit codewords and so on

9

until all 18 symbols are coded. The resulting ME code is shown in Table 1. In the next section, we will generalize the ME coding algorithm and show that this algorithm provides the optimal fixed-length code in terms of energy efficiency.

3 Memoryless ME Coding

In this section we investigate the following problem.

Given a source alphabet and source symbol probabilities, what is the optimal energy code for RF transmission?

Interestingly the solution we propose, ME Coding, is similar to well-known Huffman coding. Huffman coding is a variable-length coding technique that aims to achieve optimal compression by assigning shorter codewords to higher probability symbols. In a departure from this approach, ME Coding aims to provide optimal energy efficiency in RF transmission by minimizing average number of high bits in a codeword.

3.1 ME Coding

We start investigating the minimum energy source encoding problem with the simplest case of fixed-length, memoryless coding. First, we give a formal definition of ME Coding.

Definition: ME Coding. ME Coding is a statistical coding technique that assigns codewords with fewer high bits to source symbols with higher probabilities.

Given a source alphabet $S = \{s_1...s_q\}$ with symbol probabilities

$$P = \{P(s_1) \geq P(s_2) \geq P(s_3) \quad ... \quad P(s_{q-1}) \geq P(s_q)\},$$

ME Coding assigns codewords to source symbols such that

$$N = \{n_1 \leq n_2 \leq n_3 \quad ... \quad n_{q-1} \leq n_q\},$$

where n_i denotes the number of high bits in the codeword for symbol s_i.

3.2 ME Coding Theorem

Theorem: ME Coding Theorem. *Among all fixed-length, memoryless codes, ME Coding results in the code having the minimum average number of high bits. Hence, ME Coding results in the optimal energy code for RF transmission.*

Proof. We will now prove that ME Coding results in optimal energy codes providing the minimum average number of high bits in a codeword. ME coding requires that for any x < w we have both conditions

$$P(s_x) \geq P(s_w) \quad and \quad n_x \leq n_w. \tag{3}$$

In computing the average number of high bits using

$$n_{\mathrm{ave}} = \sum_{i=1}^{q} n_i P(s_i),$$

we have, among others, the two terms

$$Old: \quad P(s_{\mathrm{x}})n_{\mathrm{x}} + P(s_{\mathrm{w}})n_{\mathrm{w}}.$$

We interchange the codewords for s_{x} and s_{w}, and get the following terms

$$New: \quad P(s_{\mathrm{x}})n_{\mathrm{w}} + P(s_{\mathrm{w}})n_{\mathrm{x}}.$$

We subtract *Old* from *New* to obtain the change due to this reassignment:

$$\begin{aligned} New - Old =& P(s_{\mathrm{x}})(n_{\mathrm{w}} - n_{\mathrm{x}}) + P(s_{\mathrm{w}})(n_{\mathrm{x}} - n_{\mathrm{w}}) \\ =& (P(s_{\mathrm{x}}) - P(s_{\mathrm{w}}))(n_{\mathrm{w}} - n_{\mathrm{x}}) \geq 0. \end{aligned} \tag{4}$$

Due to ME Coding requirements in equation (3) this is a non-negative number: we can't decrease the average number of high bits if we interchange the codewords for any s_{x} and s_{w}. Hence, we conclude that ME Coding provides optimal energy codes with the minimum average number of high bits in a codeword.

An important point to note is that, like Huffman coding, ME Coding is most beneficial when the source symbols have a low variance probability distribution. Considering the other extreme, if the source symbols have equal probabilities, any L-bit binary code using all 2^L available codewords will perform as good as the ME code.

3.3 ME Coding Algorithm

We will now demonstrate the ME Coding algorithm. Given the source alphabet, S, the symbol probabilities, $P(s_i)$, the minimum number of points we want to transmit per second, m, and the bit period, t_{b}, we begin by determining the codeword length, L. The number of source symbols, q, requires that $2^L \geq q$; and the minimum number of source symbols to be transmitted per second requires that $L \leq 1/(mt_{\mathrm{b}})$. We choose the largest L satisfying these two constraints, so that we have more codewords with fewer high bits. Once the codeword length is chosen, the symbols can be coded.

Figure 5 illustrates the ME Coding algorithm. We start by assigning the codeword of all low bits to the most probable symbol. Next, the set of codewords with one high bit are assigned to the next set of most probable symbols. Then, the set of codewords with two high bits are

s_1	s_2- s_{L+1}	s_{L+2}...		... s_q	
$\binom{L}{0}$	$\binom{L}{1}$	$\binom{L}{2}$		$\binom{L}{L\text{-}1}$	$\binom{L}{L}$

Figure 5: ME Coding Algorithm

assigned to the following set of most probable symbols. This procedure is repeated until all q symbols have been coded. The decoding of the generated code is performed using a look-up table. In the next section we will present a technique that further improves the energy efficiency in RF transmission.

3.4 How much energy can we save?

The next question we ask is *"How much energy can we save?"* To answer this question, we will first identify a suitable, widely-used coding and modulation technique to make a performance comparison. In mobile radio communications symbols being transmitted are first transformed into binary coded decimal (BCD) format, which is illustrated in Figure 6. For a given codeword length, BCD first divides the range of symbol values into a given number of quantization levels. Then, starting with the codeword of all low bits, assigns the codewords to source symbols in ascending order. The generated binary data constitutes a bit stream that will be modulated onto a carrier. The most popular modulation techniques include frequency modulation, amplitude modulation, pulse code modulation (PCM), angle modulation, linear modulation, constant envelope modulation, and spread spectrum modulation [12].

The choice of the right modulation technique for a given application depends on the channel characteristics and the desired bit error probability. The modulation technique should satisfy the following requirements

1. The modulated signal should have enough energy to be detected at the receiver end.

2. The modulation technique should produce signals whose frequency content lies within the channel bandwidth.
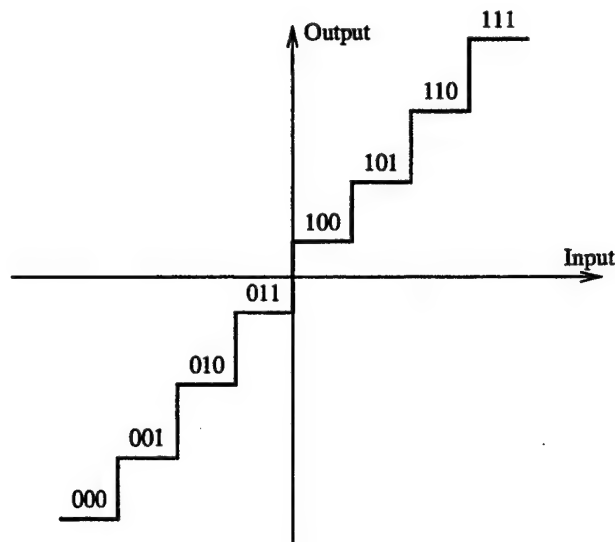
13

Figure 6: Binary Coded Decimal Values

3. The modulated waveforms should be confined in a time duration less than the bit rate to prevent inter-symbol interference.

Given these constraints, the right kind of modulation has to be chosen. Among the above listed modulation techniques, PCM constitutes a widely-used modulation technique. Hence, we will use PCM to demonstrate the amount of savings provided by ME Coding.

PCM basically produces an amplitude modulated signal to represent the binary bit stream sent from the CPU [13]. This modulation scheme was illustrated in Figure 3. The advantage of PCM is that it has the lowest bit error probability, since it has a high signal-to-noise ratio. PCM itself has various different standards determined by the Inter Range Instrumentation Group (IRIG): non-return to zero (NRZ), return to zero (RZ), biphase (Biϕ). Among these standards there are some modulation formats: level modulated (L), mark modulated (M), space modulated (S). Among these methods, the most popular used one is NRZ-L. NRZ-L basically holds the value of the binary digit at its level value for the whole bit period. This is illustrated in Figure 7. Hence, we will make our comparison against NRZ-L PCM.
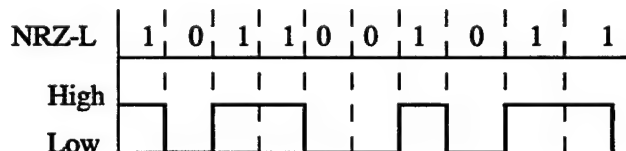

Figure 7: IRIG Standard NRZ-L PCM

ME Coding, as it was discussed earlier, is a statistical coding method. Therefore, the amount of savings obtained through ME Coding is highly-dependent on the source probability distribution. In speech coding and various other coding applications, source probabilistic models are constructed using Gaussian distributions. Thus, we will use Gaussian distributions to demonstrate when ME Coding is most beneficial. We will start with the simplest case, where we keep the codeword length constant, that is we do not tradeoff transmission rate for energy efficiency. Figure 8 shows a Gaussian probability distribution. Gaussian probability distributions have
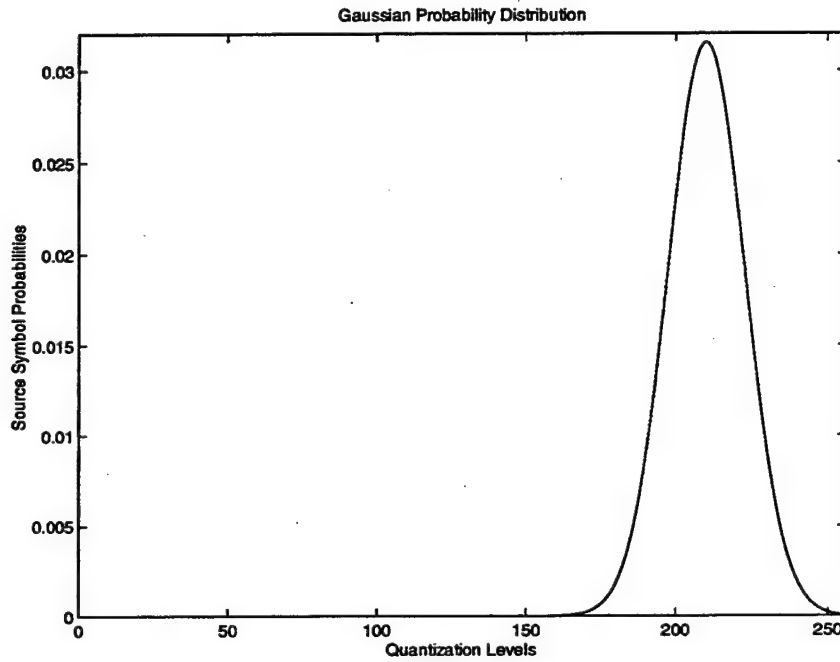

Figure 8: Source Symbol Probability Distribution

the following characteristic equation

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \; e^{-\frac{(x - m_x)^2}{2\sigma^2}} \tag{5}$$

where m_x is the mean and σ^2 is the variance of the random variable. The mean and the variance in our example are 210 and 20 respectively. For this distribution, we calculate the probability for 256 quantization levels and generate the NRZ-L PCM binary code and ME Code. We calculate the average number of high bits for each code using

$$n_{ave} = \sum_{i=1}^{q} n_i P(s_i),$$

15

and use these values to calculate the percentage power consumption reduction

$$\%\delta n_{\mathrm{ave}} = \frac{|[n_{\mathrm{ave}}]_{\mathrm{PCM}} - [n_{\mathrm{ave}}]_{\mathrm{ME}}|}{n_{\mathrm{ave-PCM}}} = 60.8\% \tag{6}$$

Figure 9 shows percentage power consumption reduction as a function of codeword length. In this figure we see that the percentage power consumption reduction for 8-bit codewords
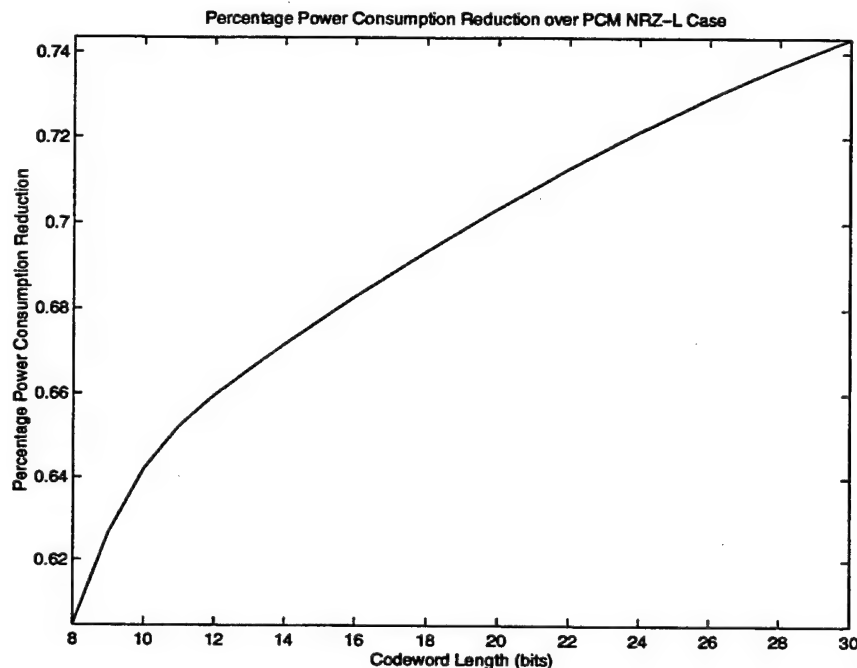

Figure 9: Percentage Power Consumption Reduction by ME Coding

using ME Coding, as it was calculated above, is 60.8%. We also see that further reduction is attained if sacrifice in transmission rate is tolerated, that is if we increase the codeword length. We see that when the codeword length is increased from 8 to 30 bits, the reduction in power consumption increases by 13.6%, resulting in a total power consumption reduction of 74.4%. The maximum improvement in power consumption performance is achieved when the codeword length is increased such that $L = q - 1$. This corresponds to a total power consumption reduction of 77.6 %. Further codeword length increase does not reduce power consumption.

Further investigation of the above example reveals some interesting points. NRZ-L PCM binary codes are constructed by mapping the ascending values of source symbols to ascending binary codeword values starting with the lowest binary value. This methodology neglects the statistical characteristics of the source symbols. Hence, the energy efficiency of the resulting code depends on the mean value of the source symbols and the variance of the source symbols.

16

Thus, the percentage reduction in power consumption via ME Coding depends in the mean and the variance of the source symbols. In the example we provide above, we obtained a 60.8% improvement in power consumption. The worst case improvement in power consumption reduction would be 0%, when the number of source symbols is equal to 2^L and the symbols are equiprobable. The best case would be 100% improvement, that is zero power consumption by ME Coding, which occurs when a certain symbol is transmitted all the time. For all other cases we will have improvements in between these two limits. The extent of improvement is determined by the variance and the mean of the source symbols. When the variance of the symbols is low and the mean of the source symbols coincides with a quantization value that has a high number of high bits in the binary equivalent, we obtain the maximum improvement. Furthermore, by increasing the codeword length, we can obtain further energy saving until $L = q - 1$.

Next, we will determine the parameters determining the maximum energy saving and derive a bound for optimal performance. Before moving onto the next section, we recommend that you to refer to Appendix B to refresh your memory with some basic concepts in information theory.

3.5   ME Coding Optimality Bound

We start investigating the optimality bound with the fixed-length, memoryless case. Derivation of the optimality bound starts with the definition of a binomial bound

$$K = \frac{1}{\left(\frac{3}{2}\right)^L} \sum_{i=1}^{q} 2^{-n_i} \leq 1. \tag{7}$$

Using equation (7), we define the numbers Q_i (pseudo probabilities):

$$Q_i = \frac{2^{-n_i}}{K\left(\frac{3}{2}\right)^L}, \tag{8}$$

where $\sum_{i=1}^{q} Q_i = 1$. The Q_i may be regarded as a probability distribution. Therefore, we can use the fundamental Gibbs inequality

$$\sum_{i=1}^{q} P(s_i) log_2 \left(\frac{Q_i}{P(s_i)}\right) \leq 0. \tag{9}$$

Upon expanding the *log* term into a sum of *log*s, we notice that one term leads to the entropy function

$$H(S) = \sum_{i=1}^{q} P(s_i) log_2 \left(\frac{1}{P(s_i)}\right) \leq \sum_{i=1}^{q} P(s_i) log_2 \left(\frac{1}{Q_i}\right).$$

17

Using equations (7) and (8) we obtain

$$H(S) \leq \sum_{i=1}^{q} P(s_i) log_2 2^{n_i} + \sum_{i=1}^{q} P(s_i) log_2 \left(\sum_{i=1}^{q} 2^{-n_i} \right),$$

$$H(S) \leq n_{\text{ave}} + log_2 \left(\sum_{i=1}^{q} 2^{-n_i} \right).$$

After some manipulation, we obtain the fundamental relationship among average number of high bits, n_{ave}, the entropy, $H(S)$, and the codeword length, L

$$2^{(H(S)-n_{\text{ave}})} \leq \sum_{i=1}^{q} 2^{-n_i}$$

where n_i is the number of high bits in the ith codeword and q is the number of source symbols. The right hand side of the inequality (3.5) is a function of the codeword length and the number of source symbols. This can be seen clearly if this expression is written in expanded form:

$$\sum_{i=1}^{q} 2^{-n_i} = \frac{1}{2^0} \binom{L}{0} + \frac{1}{2^1} \binom{L}{1} + ... + \frac{1}{2^a} \binom{L}{a} + b\frac{1}{2^{a+1}}, \tag{10}$$

where a and b are some constant integers, whose values depend on L and q.

Inequality (3.5) represents the fundamental result that we need: *Given the codeword length and the number of symbols to code, the source entropy and the codeword length supply a lower bound for the average number of high bits in a codeword.* Looking at inequality (3.5), a decrease in entropy allows a decrease in n_{ave}. Likewise, an increase in codeword length increases (10), which permits a decrease in n_{ave} in (3.5). Hence, we conclude that energy efficiency for RF transmission can be optimized either by decreasing source entropy or by increasing the codeword length. Next, we introduce a technique that further improves the energy efficiency in RF transmission.

4 Concatenation

In Section 3 we proposed the optimal energy, fixed-length, memoryless code. The next question we ask is: *Can we do better?* To answer this question we go back to Huffman coding. It is well known that coding extensions of the source alphabet (i.e., concatenations) enables Huffman coding to achieve some further compression [14]. Hence, to improve the energy efficiency, we look into concatenations of ME Coding.

4.1 What is concatenation?

Let us first define what the concatenation technique is. Concatenation is a technique that

1. Transforms a given source alphabet, S, with symbols probabilities, $P(s_i)$, into a new source alphabet, S^p, with symbol probabilities, $P(s^q)$,

2. Codes the new source alphabet into codewords of length pL using the ME Coding algorithm.

Hence, concatenation intertwines two separate mechanisms: *(i)* probability distribution alteration, and *(ii)* available number of high bits variation. The relation between the two mechanisms is judiciously utilized to attain further performance improvements.

For the simplest case we assume that the symbols are statistically independent and calculate the corresponding symbol probabilities using $P(s^q) = P(s_i)P(s_j)$. We should note that this case is not only the simplest case, but also the worst case. If the symbols are not statistically independent, that is if the mutual information is not zero, then we can achieve further power consumption reduction by utilizing this mutual information.

4.2 Concatenation Algorithm

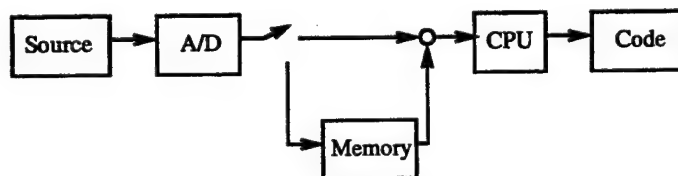The coding block diagram in concatenated ME Coding is illustrated in Figure 10. Looking at

Figure 10: Concatenation block diagram

Figure 10, we see that concatenation algorithm involves the following four steps

19

1. A source symbol is sensed, converted into digital form, and stored in memory.

2. A second symbol is sensed, digitized, and stored in memory.

3. Step two is repeated until the number of symbols stored in memory is equal to the desired concatenation order.

4. Once the concatenation order is reached, a codeword for the concatenation of the symbols in the memory is constructed according to a look-up table that is generated via ME Coding algorithm.

Figure 11 illustrates the resulting code for a second order concatenation. Here, instead of
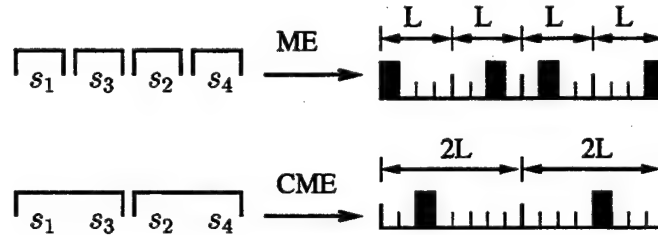

Figure 11: Concatenation mechanism

sampling each source symbol and encoding one-by-one using codewords of length L, we sample the first symbol, store it in memory, sample the second symbol, and encode the concatenation of the two symbols using codewords of length $2L$.

An immediate point to note is that, unlike ME Coding, concatenation requires memory. One drawback of using memory is that it introduces delay, which might be undesirable in some applications. A second issue is that concatenation becomes impractical as the concatenation order increases. Since the encoding is in the form of a look-up table, as the concatenation order increases, the source alphabet becomes very large resulting in impractical memory requirements and computation complexity.

4.3   Understanding concatenation

Now we ask the question *How does concatenation improve the power consumption performance of ME Coding?*. We answer this question using two approaches. We start with a low level approach, where we investigate the concatenation mechanism by analyzing the equation for average number of high bits ((2)) on a term by term basis. Then, we use a high level approach and analyze the relation between average number of high bits, variance, and symbol probabilities as a function of concatenation order.

20

4.3.1 Low level approach

In this section, we will use an example to explicate the concatenation mechanism and then generalize our result. Consider the problem of coding $q = 4$ symbols with probabilities $P_1 \geq P_2 \geq P_3 \geq P_4$ into codewords of length $L = 2$. We start by generating the ME Code for this source alphabet: [00 01 10 11], where the codewords to the four symbols have zero, one, one, and two high bits. Thus, the average number of high bits for this code becomes

$$n_{\text{ave}} = (0P_1) + (1P_2) + (1P_3) + (2P_4). \tag{11}$$

Next we decompose the source symbol probabilities as

$$P_i = P_i(P_1 + P_2 + P_3 + P_4)$$

and rewrite each term in (11) according to this decomposition

$$(0P_1) = P_1((0P_1) + (0P_2) + (0P_3) + (0P_4)) \tag{12a}$$
$$(1P_2) = P_2((1P_1) + (1P_2) + (1P_3) + (1P_4)) \tag{12b}$$
$$(1P_3) = P_3((1P_1) + (1P_2) + (1P_3) + (1P_4)) \tag{12c}$$
$$(2P_4) = P_4((2P_1) + (2P_2) + (2P_3) + (2P_4)) \tag{12d}$$

Hence, the average number of high bits for the ME Code becomes the sum of these four expressions.

Correspondingly, we construct the first concatenation of this source alphabet and construct the accompanying ME Code. For the first concatenation we have $q = 16$ symbols, codeword length $L = 4$, and symbols with probabilities $P_i P_j$. We again construct the available codewords for a code of length $L = 4$ and calculate the number of high bits in each codeword. The vector of number of high bits in each codeword becomes

$$n = [0\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 4].$$

Keep in mind that by coding the first order concatenation of the source alphabet, we will be sending two symbols at a time. Hence, the number of high bits per symbol becomes

$$\frac{n}{2} = [0\ 0.5\ 0.5\ 0.5\ 0.5\ 1\ 1\ 1\ 1\ 1\ 1\ 1.5\ 1.5\ 1.5\ 1.5\ 2]. \tag{13}$$

The corresponding average number of high bits for the concatenated ME Code becomes

$$n_{\text{ave}} = \sum_{i=1}^{16} n_i P_i \tag{14}$$

21

The terms in this equation can be grouped and factorized as

$$P_1((aP_1) + (bP_2) + (cP_3) + (dP_4)) \tag{15a}$$

$$P_2((eP_1) + (fP_2) + (gP_3) + (hP_4)) \tag{15b}$$

$$P_3((kP_1) + (mP_2) + (oP_3) + (rP_4)) \tag{15c}$$

$$P_4((sP_1) + (tP_2) + (uP_3) + (vP_4)) \tag{15d}$$

where $[a \quad ... \quad v]$ are the number of high bits per source symbol for the available codewords. The choice of the mapping between $[a \quad ... \quad v]$ and $[0 \; 0.5 \; 0.5 \; 0.5 \; ... \; 1.5 \; 2]$ is left to the code designer.

At this point we realize how concatenation achieves further improvement in power consumption reduction. Looking at (12), we see that the coefficients in each term for the equations are coupled. That is each term in each equation is assigned the same number of high bits. However, this is not the case in concatenation. Looking at (15), we see that the coefficients are independently assignable. Hence, we can assign the available number of high bits to each symbol to further reduce the total power consumption reduction. The constraint that we still have is that the values we assign are limited to $[0 \; 0.5 \; 0.50.5 \; ... \; 1.5 \; 2]$. Hence, some of the terms in (14) will decrease and some will increase, but the overall effect will enable power consumption reduction. A similar argument can be made when coding sources, where $q \leq 2^L$. For higher order concatenations the sub-probabilities are decomposed to sub-sub-probabilities and additional control over these probabilities is obtained a similar decoupling mechanism.

Result. *Concatenation decomposes the source symbol probabilities into a new set of sub-probabilities, thereby decoupling these sub-probabilities. Since, each coefficient in (15) is independently accessible, further power consumption reduction becomes possible through proper source coding.*

4.3.2 High level approach

As an alternative approach, we can investigate the concatenation problem using the variance of a random variable. Hence, let us start by defining what variance is.

Variance. The variance of a random variable n_i, denoted by $var(n)$, is the mean squared deviation of n_i from its average value. The equation for $var(n)$ is given as

$$var(n) = \sum_{i=1}^{q} (n_i^2 - n_{ave}^2) P(s_i). \tag{16}$$

where n_{ave} is the average number of high bits, n_i is the number of high bits in the ith codeword, and $P(s_i)$ is the probability of ith source symbol. In this equation, $var(n)$, n_i, n_{ave}^2, and $P(s_i)$ all change as functions of the concatenation order. Rearranging equation 16, we obtain the

equation for the average number of high bits as a function of the concatenation order, the variance, the symbol probabilities, and the number of high bits in each codeword

$$n_{\text{ave}}(p) = \frac{\sqrt[2]{\sum_{i=1}^{q^p} n_i(p)^2 P(s_i(p)) - var(n,p)}}{p}. \tag{17}$$

When coding higher order concatenations of a source alphabet, the sum term $\sum_{i=1}^{q^p} n_i(p)^2 P(s_i(p))$ and the variance $var(n,p)$ both increase. However, the relation between the two terms is such that the average number of high bits increases. This is illustrated in Figure 12. In Figure 12
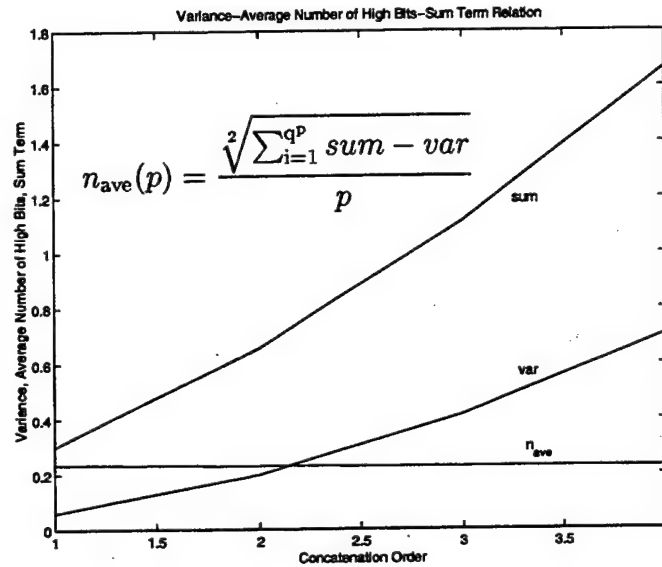

Figure 12: Variance–Average Number of High Bits–Sum Term Relation

we see that both the sum term and the variance increases. However, square root of the difference of the two terms is not increasing as fast as the concatenation order, p. Therefore, we obtain power consumption reduction through concatenation. We conclude that by increasing the variability of the source symbols, concatenation achieves some further power consumption reduction.

So far we have seen that the following techniques can improve the power consumption performance of ME Coding

1. Increasing codeword length L

2. Decreasing source entropy H(S)

3. Increasing concatenation order p.

Next we will ask the question *How much more energy can we save?*

23

4.4 How much more energy can we save?

For a given number of source symbols to code, the amount of energy savings obtained through concatenation depends on the source probabilistic distribution and the initial codeword length. Figure 13 is a representative plot illustrating the amount of savings obtained when coding a 16 symbol source having a Gaussian probabilistic distribution. After qu   zation and generating
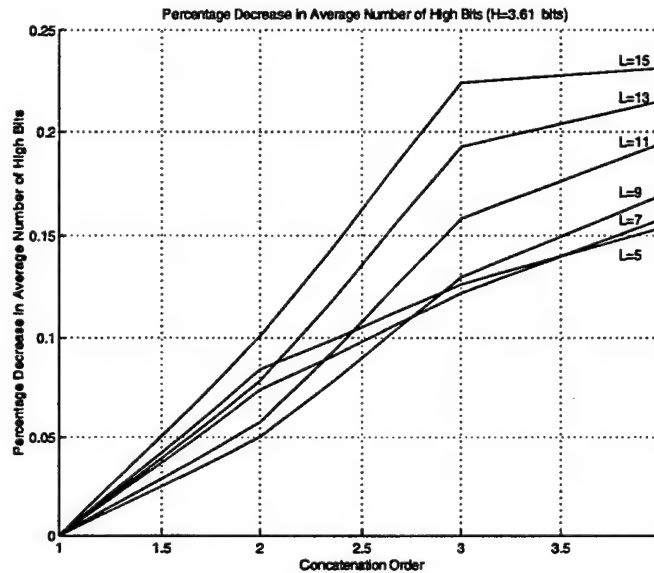

Figure 13: Energy saving through concatenation for a high-entropy source

the probability values for the source symbols, we find that the source entropy is $H = 3.61$ bits. The shortest codeword length that can be used is 4 bits long ($L \geq log q$). Hence, we start with $L = 4$ and generate concatenations of the ME Code to find the amount savings that can be obtained. Analyzing the figure we note the following points. For this high entropy source, the total percentage power consumption reduction increases as the concatenation order increases. The amount of total percentage power consumption reduction changes with the initial codeword length. The total percentage power consumption reduction through concatenation is on the order of 30%.

Let us now investigate the percentage power consumption reduction for a low entropy source, $H = 1.45$ bits. Figure 14 shows the amount of power consumption reduction obtained for this source. In this case we again see that the total percentage power consumption reduction increases with concatenation order. However, the main point that we want to illustrate is that for this low entropy source, the total percentage power consumption reduction through concatenation is only on the order of 3% to 8% depending on the initial codeword length. This kind of behavior is a characteristic of low entropy sources. Hence, we conclude that
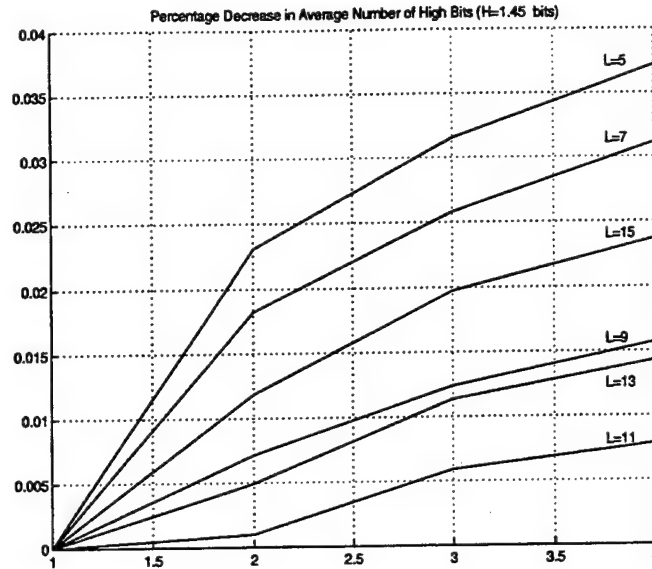
24

Figure 14: Energy saving through concatenation for a low-entropy source

concatenation is most useful when coding high entropy sources. In low entropy sources, ME Coding solely takes care of the attainable power consumption reduction and the impact of concatenation is comparably small.

Let us summarize the insight we gained about concatenation through the examples, analysis and simulations (Due to space limitations, we can't include the other examples that we used to derive the following insights).

- For a given number of symbols to code, the total percentage power consumption reduction through concatenation depends on the source entropy and the initial codeword length.

- The total percentage power consumption reduction through concatenation generically increases with increasing concatenation order.

- The highest percentage power consumption reduction generally occurs in the first and second concatenations depending on the initial codeword length and decreases for higher order concatenations.

- The highest total power consumption reduction by concatenation occurs for high entropy sources. The amount of savings is up to 30%.

- The lowest total power consumption reduction by concatenation occurs for low entropy source. The amount of savings is around 5%.

In the next section, we will derive an optimality bound for concatenation performance.

25

4.5   Concatenated ME Coding Optimality Bound

Optimality bound for concatenation is obtained from the bound in (3.5). The pth concatenation of a source increases the number of source symbols from q to q^p. If the source symbols are statistically independent, then $H(S^p) = pH(S)$. Hence, making these substitutions into (3.5) and simplifying gives the following optimality bound for concatenation

$$2^{(H(S) - \frac{n_{\text{ave}}}{p})} \leq \sqrt[p]{\sum_{i=1}^{q^p} 2^{-n_i}}. \tag{18}$$

By varying the concatenation order, the right hand side in (18) can be increased, permitting a decrease in (n_{ave}/p). This behavior is shown in Figure 15. The increasing behavior of this
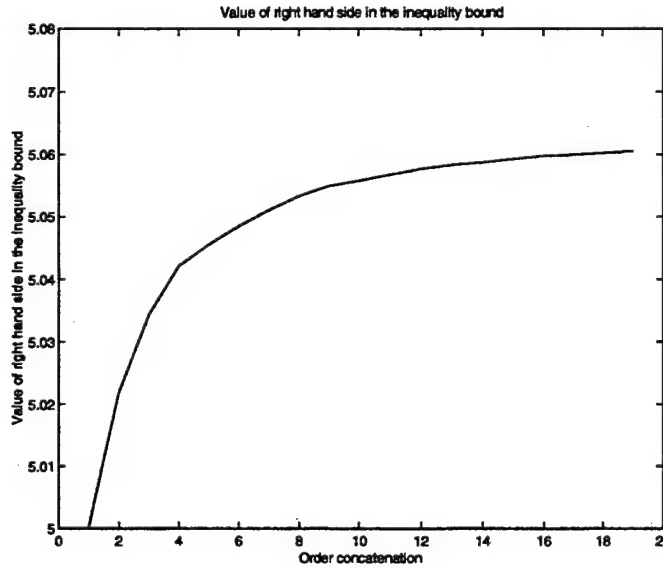

Figure 15: Bound for Concatenation Power Consumption Performance

function means that as the right hand side in (18) increases, a further decrease in n_{ave} becomes possible.

5 Discussion

5.1 Specific Contributions

Traditional source coding aims to optimize data compression for maximum transmission rate. In a recognition of the increased technological demand for energy efficient wireless communication techniques this research reformulates the source coding problem to achieve optimum energy efficiency. This reformulation results in the following significant contributions.

- We formulate the power consumption equation for a standard RF transmitter and propose various strategies for optimizing the power consumption performance.

- We reformulate the source coding problem for energy efficient wireless transmission.

- We propose a novel coding algorithm, ME coding, that provides optimal energy codes given the source statistics.

- We prove optimality of ME Coding.

- We compare the power consumption performance of ME Coding against that of PCM.

- We identify the parameters determining optimality and derive an optimality bound.

- We elucidate the concatenation mechanism and present further insight into the concatenation behavior.

- We propose a bound on optimal concatenation performance.

5.2 Results and Conclusions

The analysis and simulations performed in this project lead to the following results:

5.2.1 Results on ME Coding

1. Among all fixed-length, memoryless codes, ME Coding results in the code having the minimum average number of high bits. Hence, ME Coding results in the optimal energy code for RF transmission.

2. The power consumption performance of ME Coding depends on the source entropy and the codeword length.

3. For a generic Gaussian source probability distribution and a codeword length of $L = log_2\text{q}$, ME Coding improves the energy efficiency of NRZ-L PCM by <u>60.8%</u>. This figure may increase or decrease depending on the source probability distribution and the codeword length. We should note that this improvement is obtained without any tradeoffs for transmission rate.

4. For the same Gaussian probability distribution the energy efficiency can further be improved to <u>74.4%</u> by making some sacrifice in transmission rate (increasing codeword length). In the limit this value converges to <u>77.6%</u>. Again this figure of savings is source probability distribution dependent.

5.2.2 Results on Concatenation

1. Coding concatenations of the source symbols provides additional energy saving.

2. For a given number of symbols, the percentage power consumption reduction through concatenation depends on the source entropy, the initial codeword length, and the number of source symbols to code.

3. The total power consumption reducti_n through concatenation generically increases with increasing concatenation order.

4. The highest percentage power consumption reduction generally occurs in the first and second concatenations depending on the initial codeword length. It decreases for higher order concatenations.

5. The highest total power consumption reduction by concatenation occurs for high entropy sources. The amount of saving is up to <u>30%</u>.

6. The lowest total power consumption reduction by concatenation occurs for low entropy sources. The amount of saving is around <u>5%</u>. This is due to the fact that ME Coding alone provides the majority of energy savings for low entropy sources.

The results provided above show that ME Coding and Concatenated ME Coding can provide significant performance improvements over the existing coding and modulation algorithms. Future efforts in developing ME Coding will be in error detection and correction.

5.3 Future Directions

Future efforts in developing ME Coding will concentrate on error recovery and implementation. First, energy performance of existing error control strategies will be compared. Then, a novel

error control technique will be developed and merged into ME Coding. Finally, ME Coding algorithm will be implemented in a codec and will be used in the ring sensor and various other wireless communications applications.

References

[1] T. Imielinski, M. Gupta, and S. Peyyeti. Energy efficient data filtering and communication in mobile wireless computing. In *Proceedings of the Second USENIX Symposium on Mobile and Location-Independent Computing*, pages 109–119, April 1995.

[2] W. Mangione-Smith, P. S. Ghang, S. Nazareth, P. Lettieri, W. Boring, and R. Jain. A low power architecture for wireless multimedia systems: Lessons learned from building a power hog. In *Proceedings of the 1996 International Symposium on Low Power Electronics and Design*, pages 23–28, 1996.

[3] I. Chlamtac, C. Petrioli, and J. Redi. An energy-conserving access protocol for wireless communication. In *Proceedings of ICC'97 - International Conference on Communications*, volume 2, pages 1059–1061. IEEE, 1997.

[4] T. H. Meng. A wireless portable video-on-demand system. In *Proceedings of the Eleventh International Conference on VLSI Design*, pages 4–9. IEEE, 1997.

[5] M. B. Srivastava, A. P. Chandrakasan, and R. W. Brodersen. Predictive system shutdown and other architectural techniques for energy efficient programmable computation. *IEEE Transactions on VLSI Systems*, 4(1):42–55, March 1996.

[6] C.-L. Su, C.-Y. Tsui, and A. M. Despain. Saving power in the control path of embedded microprocessors. *IEEE Design and Test of Computers*, 11(4):24–31, 1994.

[7] T.-H. Lan and A.-H. Tewfik. Adaptive low power multimedia wireless communications. In *Proceedings of the 1997 IEEE First Workshop on Multimedia Signal Processing*, pages 377–382. IEEE, 1997.

[8] B.-H. Yang, S. Rhee, and H. H. Asada. A twenty-four hour tele-nursing system using a ring sensor. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, volume 1, pages 387–392. IEEE, 1998.

[9] D. W. Faulkner. PCM signal coding. Patent Number 5,062,152, October 1991.

[10] J. R. Smith. *Modern Communication Circuits*. McGraw-Hill, 1998.

[11] T. E. Price. *Analog Electronics: an integrated PSpice approach*. Prenctice Hall, 1997.

[12] T. S. Rappaport. *Wireless Communications Principles*. Prentice Hall, 1996.

[13] B. Waggener. *Pulse Code Modulation Techniques*. Salomon Press, 1995.

[14] R. W. Hamming. *Coding and Information Theory*. Prentice-Hall, 1986.

[15] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[16] J. G. Proakis. *Digital Communications*. McGraw-Hill, 1995.

Appendix A: Source Coding for Digital Communications

Figure 16 illustrates the functional diagram and the basic elements of a digital communication system [15, 16]. In this diagram, the source output may be an analog signal or a digital signal.
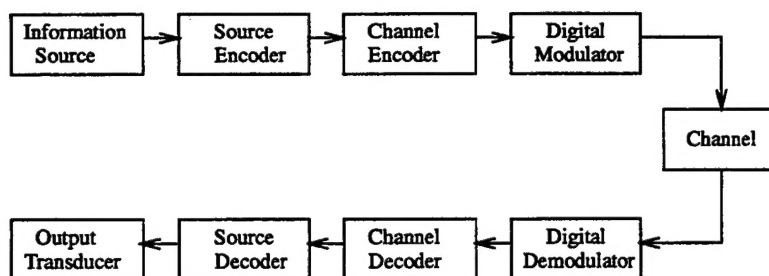

Figure 16: Block diagram of a digital communication system

In a digital communication system, the messages produced by the source are converted into a sequence of binary digits. We ideally seek to represent the source output in an efficient manner according to our preference. The process of efficiently converting the output of either an analog or digital source into a sequence of binary digits is called *source encoding* or *data compression*.

The sequence of binary digits from the source encoder, which we call the information sequence, is passed to the channel encoder. The purpose of the channel encoder is to introduce, in a controlled manner, some redundancy into the binary information sequence. This redundancy is used at the receiver to overcome the effects of noise and interference encountered in the transmission of signal through the channel.

The binary sequence at the output of the channel encoder is passed to the digital modulator, which serves as the interface to the communications channel. The purpose of the modulator is to map the binary information sequence into signal waveforms. The communication channel is the physical medium that is used to send the signal from the transmitter to the receiver. In wireless transmission, the channel is the atmosphere. Whatever the physical medium is, the transmitted signal is corrupted in a random manner by a variety of possible mechanisms, such as thermal noise, man-made noise, automobile ignition noise, and electrical lightning discharges. At the receiving end, the digital demodulator processes the channel-corrupted transmitted waveform and reduces the waveform to a sequence of numbers that represent estimates of the transmitted data symbols. This sequence of numbers is passed to the channel decoder that attempts to reconstruct the original information sequence from knowledge of the code used by the channel encoder and the redundancy contained in the received data.

Appendix B: Information and Entropy

Information

Suppose that we have the source alphabet of q symbols $S = \{s_1...s_q\}$ each with its probability $P = \{P(s_1), P(s_2), P(s_3) \quad ... \quad P(s_{q-1}), P(s_q)\}$ [14]. When we receive one of these symbols, how much information do we get? For example, if $P(s_1) = 1$, then there is no "surprise", no information, since you know what the message must be. On the other hand, if the probabilities are all very different, then when a symbol with low probability arrives, you feel more surprised, get more information, than when a symbol with higher probability arrives. Thus information is somewhat related to the inverse of symbol probability.

To construct the information function, $I(P(s_i))$, we assume three things

1. $I(P(s_i)) \geq 0$ (a real nonnegative measure)

2. $I(P(s_i)P(s_j)) = I(P(s_i)) + I(P(s_j))$ for independent events (additive)

3. $I(P)$ is a continuous function of P.

The second of these conditions is known as the *Cauchy functional equation* for the function $I(P)$, meaning that it serves to define $I(P)$. We recognize that the *log* function obeys all the three conditions, hence, the equation for information function becomes

$$I(P) = log\frac{1}{P}. \tag{19}$$

The next question we ask is *"what base of log system shall we use?"* It is simply a matter of convention since any set of *logs* is proportional to any other set. It is convenient to use the base 2 *logs*; the resulting unit of information is called a *bit*.

Entropy

If we get $I(s_i)$ units of information when we receive the symbol s_i, how much do we get on the average?

Since $P(s_i)$ is the probability of getting the information $I(s_i)$, then on the average we get for each symbol s_i

$$P(s_i)I(s_i) = P(s_i)log\frac{1}{P(s_i)}. \tag{20}$$

From this it follows that on the average, over the whole alphabet of symbols s_i, we will get

$$H(S) = \sum_{i=1}^{q} P(s_i)I(s_i) = P(s_i)log\frac{1}{P(s_i)}. \tag{21}$$

where $H(S)$ is called the *entropy* of the source S.

The entropy function involves only the distribution of the probabilities – it is a function of a probability distribution and does not involve the symbol values.